Wastewater Surveillance of SARS-CoV-2 for Predicting COVID-19 Cases in Minnesota

Kah Meng Soh & King Yiu Suen
University of Minnesota

Abstract

Wastewater-based epidemiology provides an approach for assessing the prevalence of COVID-19 in a sewer service area. In this study, SARS-CoV-2 RNA was detected in 40 wastewater treatment plants of varying sizes and served populations across the state of Minnesota during 2022. Various linear regression models were investigated to predict the weekly case count from SARS-CoV-2 RNA concentrations under various transformation and normalization methods. It is found that the relationship between COVID-19 incidence and SARS-CoV-2 RNA in wastewater may be treatment plant-specific. Including the vaccination rate in the model may be helpful but the results are not very robust over different forecast horizons. The case count of the previous week tends to be a stronger predictor than SARS-CoV-2 RNA concentration

Contents

1 Introduction

Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has infected millions of people across the globe and resulted in significant health and economic impacts. Consequently, methods for detecting and tracking this infectious disease at the community level are urgently needed.

Mass community testing is costly and the demand for tests frequently exceeds the capacity of testing facilities (Barasa et al., 2020), not to mention that not everyone has access to testing due to economic, geographic, or social restrictions. Furthermore, test results are a lagging indicator of the pandemic's progression, because testing is usually prompted by symptoms, which may take 2 weeks to show up after infection (Lauer et al., 2020). Thus, delays may occur between the appearance of symptoms, testing and the reporting of test results (Peccia et al., 2020). Finally, it is estimated that as many as 45% of COVID-19 cases are asymptomatic (Li et al., 2020; Nishiura et al., 2020; Oran and Topol, 2020; Post et al., 2020). Considering that people only seek medical attention and undergo diagnostic testing if they are symptomatic, the number of confirmed clinical cases may grossly underestimate the prevalence of the disease.

Wastewater-based epidemiology (WBE) is an emerging method of monitoring trends of the virus in communities. In WBE, wastewater is sampled from wastewater treatment plants (WWTPs) and is tested for signatures of viruses excreted via feces. The presence of viruses in wastewater samples informs the potential of viral outbreaks in the communities served by those plants. WBE has been successfully employed as a surveillance tool for diseases such as SARS, hepatitis A, and polio (Hellmér et al., 2014; Manor et al., 1999; Ye et al., 2016). With regard to SARS-CoV-2, viral particles are reported to be shed in feces from infected individuals even if they are asymptomatic (Chan et al., 2021; Chen et al., 2020; Cheung et al., 2020; Parasa et al., 2020; Wong et al., 2020). Recent studies have shown that WBE is able to predict COVID-19 prevalence even earlier than clinical case data (Peccia et al., 2020; Ahmed et al., 2020; Arora et al., 2020; Randazzo et al., 2020), supporting the idea that WBE can be used as an early warning system to identify disease hotspots.

Estimating the SARS-CoV-2 RNA concentrations in wastewater (gene copies per litre) is complicated, as the dilution and fecal strength in the wastewater may vary between sampling dates. It has been recommended to multiply the viral concentration in wastewater by the

flow of the sampled location (the volume of wastewater that passed through the location in a day) to obtain the viral concentrations in gene copies per day, and account for changes in sanitary sewer contributions (Hasan et al., 2021; Weidhaas et al., 2021). However, the flow rate is not stable and is impacted by many factors such as rainstorms. Normalizing SARS-CoV-2 RNA concentrations by indicators of human fecal waste is also common, because feces in wastewater can have variable levels of SARS-CoV-2 depending upon the amount of water used per toilet flush or body washing (Zhan et al., 2022). The contribution of SARS-CoV-2 from human sourced water can then be estimated by dividing the measured SARS-CoV-2 concentration by the concentration of the human waste indicator (Zhan et al., 2022). A typically examined fecal marker is Pepper Mild Mottle Virus (PMMoV) (Maal-Bared et al., 2023; Zhan et al., 2022). Previous studies have shown that PMMoV is the most abundant RNA virus in human feces and it is shed in large quantities in wastewater (Hamza et al., 2019; Kitajima et al., 2014, 2018; Rosario et al., 2009; Zhang et al., 2006). It is also highly stable in wastewater, and its concentrations showed little seasonal variation (Kitajima et al., 2014, 2018).

The main objective of this study is to develop predictive models to predict the number of COVID-19 cases using wastewater samples from 40 WWTPs in Minnesota, USA from March 2022 to October 2022. In particular, the current study attempts to answer the following research questions:

1. What is the best way to incorporate SARS-CoV-2 concentrations into the model? More specifically,
    a. How should the virus concentrations be normalized?
    b. Which SARS-CoV-2 gene should be used?
    c. Does including lagged virus concentrations in the model improve the predictive performance?
    d. Does a $\log_{10}$ transformation of the variables improve the predictive performance?
2. Is it possible to develop one model for all WWTPs or is it necessary to develop one model for each WWTP?
3. Is the vaccination rate a useful predictor?
4. Is the lagged case count a more useful predictor than virus concentrations?
5. How will the predictive performance be affected if the forecast horizon is increased?

All statistical analyses were performed using R version 4.2.1 (R Core Team, 2022).

2 Data Description

Wastewater samples were collected from 40 WWTPs across the state of Minnesota. These 40 WWTPs represent a broad sampling of the Minnesota population serving a total of 191 zip codes and a population of 3,825,269 people, which is approximately 67% of the total population of the state. The data collection period varied depending on the WWTP, but was typically between March 2022 and October 2022. The wasterwater samples were collected two days per week by each WWTP. For data analysis purposes, a weekly level average of measurements was used. The virus concentrations of three SARS-CoV-2 target genes, nucleocapsid (N), spike (S), and ORF1ab (O) proteins, were measured in the wastewater samples. Figure 1 displays the concentrations of N, S and O in three WWTPs, Little Falls, Northfield and Twin Cities. Out of the 1,213 samples collected, one was reported to have zero concentration of N, S and O. It was removed in further analyses. Among the remaining samples, 424 were reported to have zero virus concentration of S. Since S had too many zero values, this variable was not used in any further analyses. The Pearson correlation coefficient between the virus concentrations of N and S was 0.95.

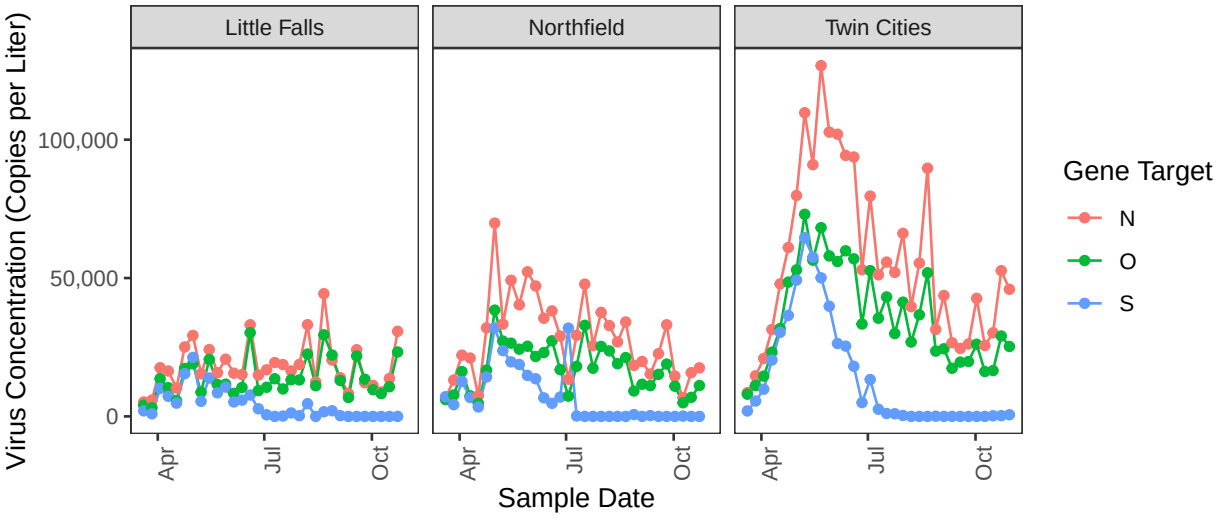

Figure 1: Virus concentrations in Little Falls, Northfield and Twin Cities

The concentrations of a human fecal marker, PMMoV, were also determined in each wastewater sample. There were a few outliers in this variable for unknown reasons. PMMoV concentrations below the 5th percentile were replaced the 5th percentile value, and those above the 95th percentile were replaced with the 95th percentile value. The influent flow rate was provided by the participating WWTPs.

The weekly number of new infections for each WWTP service area were obtained from the

Minnesota Department of Health (MDH). To allow a $\log_{10}$ transformation, any zero case counts were replaced with ones. The MDH also provided the number of people who have received at least one dose of COVID-19 vaccine, and the number of people with completed vaccine series (people who are fully vaccinated) in the areas served by each WWTP over time. The complete series could be one, two, or three doses depending on the person's age and which vaccine they received. It does not include booster doses.

Table 1 provides some descriptions of the participating WWTPs, including the sampling period, the number of samples collected, the size of the population served, as well as the mean and the standard deviations of weekly case count and flow.

3 Statistical Analysis

3.1 Use of SARS-CoV-2 Concentrations

Denote C_t as the COVID-19 case count at time t, and W_t as the SARS-CoV-2 concentrations (either O or N) measured in a wastewater sample at time t. In this section, I compared different linear regression models for predicting C_t from W_t.

As mentioned in section 1, virus concentrations are commonly normalized by either flow:

$$W_t^{\text{Flow}} = W_t \cdot \text{Flow}_t \tag{1}$$

or a fecal marker such as PMMoV:

$$W_t^{\text{PMMoV}} = W_t / \text{PMMoV}_t \tag{2}$$

However, there have been contradictory findings on whether normalization of virus concentration can improve correlations with cases (Duvallet et al., 2022; Feng et al., 2021; Maal-Bared et al., 2023). Moreover, no study has examined using both flow and PMMoV to normalize the virus concentration:

$$W_t^{\text{Flow \& PMMoV}} = W_t \cdot \text{Flow}_t / \text{PMMoV}_t \tag{3}$$

It would be interesting to compare these three normalization approaches. Another question of interest concerns the concentrations of which SARS-CoV-2 gene should be used. Since N and O are highly correlated, it is expected that they will result in similar performance. It is also of interest to know whether adding lagged virus concentrations (e.g., W_{t-1} and W_{t-2}) in

Table 1: Descriptions of participating WWTPs

WWTP	Code	No. of Samples	Population	Average Weekly Case Count
Albert Lea	AL	32	20957	27.59
Alexandria	AX	32	25689	38.44
Bemidji	BI	24	32741	33.67
Blooming Prairie	BP	18	3589	2.67
Blue Lake	BL	33	397299	617.18
Cambridge	CB	31	15529	17.94
Chisholm	CM	25	5787	7.20
Eagles Point	EP	33	113786	192.97
East Bethel	EB	32	14880	15.00
Elk River	ER	31	94119	125.42
Empire	EM	33	176441	257.88
Fergus Falls	FF	31	18799	22.68
Glacial Lakes SSWD	GL	29	5574	5.83
Glencoe	GC	15	8280	12.47
Hastings	HS	33	30027	34.67
Hinckley	HY	30	5483	6.23
Hutchinson	HT	26	17829	32.04
International Falls	IF	32	9618	20.06
Lafayette	LY	28	844	0.93
Lanesboro	LB	29	1849	1.97
Le Sueur	LS	32	8237	7.19
Little Falls	LF	32	14615	23.72
Mankato	MK	32	63489	77.88
Marshall	ML	32	15550	17.62
Moorhead	MH	27	42602	66.22
Mora	MR	32	9872	14.16
New Prague	NP	32	12557	12.69
North Branch	NB	32	16827	21.53
Northfield	NF	32	27040	33.38
Rochester	RC	32	123041	273.16
Rogers	CR	32	14573	18.81
Saint Cloud	SD	31	130901	245.39
Seneca	SN	33	243338	390.85
St. Croix Valley	SV	33	38958	59.39
Thief River Falls	TR	32	12971	17.16
Twin Cities	TC	33	1854291	2959.61
Western Lake Superior Sanitary District	WL	33	124459	222.39
Willmar	WM	32	23697	31.75
Winona	WA	30	34644	59.60
Worthington	WT	32	14487	17.03

the model will improve the predictive performance. Finally, it is prevalent in the wastewater literature to use a $\log_{10}$ transformation on the variables to meet assumptions for parametric analysis (Farkas et al., 2022; Feng et al., 2021). It would be useful to know whether a $\log_{10}$ transformation affects the predictive performance.

To summarize, I varied the following factors:

1. Normalization of the virus concentrations (unnormalized, flow, PMMoV or both)
2. SARS-CoV-2 gene to use (N or O)
3. The number of lagged values for the virus concentrations (0, 1 or 2)
4. Whether a $\log_{10}$ transformation was used on the dependent and independent variables

The four factors were fully crossed, resulting in a total of $4 \cdot 2 \cdot 3 \cdot 2 = 48$ conditions. The models were fitted separately for each WWTP in each condition, since it is unclear whether we can use one model for all WWTPs (I will explore this question in the next section). To allow for comparisons between WWTPs with different sizes of population served, I divided the case count and virus concentrations by the population size. This ensures that the prediction errors for all WWTPs are theoretically on the same scale. Observations with missing values due to the creation of lagged variables were removed.

To assess the model performance, a leave-one-out cross-validation (LOOCV) was used. The models were trained on $n - 1$ observations and validated on the remaining one observation, where n is the sample size. The procedure was repeated n times with each of the n observations used exactly once for validation. The average of the n prediction errors obtained was computed for model comparison. The evaluation metric was the root mean squared errors (RMSE) between the predicted value and the actual value. If the dependent variable was $\log_{10}$-transformed, the transformation was reversed to obtain the predictions on the original scale before the computation of RMSE.

Table 2 displays the means and standard deviations of the cross-validated RMSE of the linear regression models under different conditions across WWTPs. For ease of presentation, the RMSE were multiplied by 100. When the raw data were used, normalizing by flow slightly reduced the RMSE, whereas normalizing by PMMoV or by both flow and PMMoV led to an increase in RMSE. When a $\log_{10}$ transformation was applied on the variables, both the mean and standard deviation of RMSE tended to decrease, especially when the virus concentrations were normalized by PMMoV. The RMSE for O was slightly lower than N, but was in generally very similar. This can be explained by the high correlation between them. Including extra lags generally worsened the predictive performance.

Table 2: Means and standard deviations of RMSE across WWTPs in all 48 conditions

Gene	Normalization	Scale	Lag	Mean	SD	Gene	Normalization	Scale	Lag	Mean	SD
N	Unnormalized	Raw	0	0.06	0.02	O	Unnormalized	Raw	0	0.06	0.02
N	Unnormalized	Raw	1	0.07	0.02	O	Unnormalized	Raw	1	0.07	0.03
N	Unnormalized	Raw	2	0.07	0.03	O	Unnormalized	Raw	2	0.07	0.03
N	Unnormalized	Log	0	0.06	0.02	O	Unnormalized	Log	0	0.06	0.02
N	Unnormalized	Log	1	0.07	0.02	O	Unnormalized	Log	1	0.06	0.02
N	Unnormalized	Log	2	0.07	0.02	O	Unnormalized	Log	2	0.06	0.02
N	Flow	Raw	0	0.06	0.02	O	Flow	Raw	0	0.06	0.02
N	Flow	Raw	1	0.06	0.02	O	Flow	Raw	1	0.06	0.02
N	Flow	Raw	2	0.07	0.02	O	Flow	Raw	2	0.07	0.02
N	Flow	Log	0	0.06	0.02	O	Flow	Log	0	0.06	0.02
N	Flow	Log	1	0.06	0.02	O	Flow	Log	1	0.06	0.02
N	Flow	Log	2	0.07	0.02	O	Flow	Log	2	0.06	0.02
N	PMMoV	Raw	0	0.08	0.04	O	PMMoV	Raw	0	0.08	0.04
N	PMMoV	Raw	1	0.09	0.05	O	PMMoV	Raw	1	0.09	0.05
N	PMMoV	Raw	2	0.11	0.08	O	PMMoV	Raw	2	0.12	0.09
N	PMMoV	Log	0	0.06	0.02	O	PMMoV	Log	0	0.06	0.02
N	PMMoV	Log	1	0.06	0.03	O	PMMoV	Log	1	0.06	0.03
N	PMMoV	Log	2	0.07	0.03	O	PMMoV	Log	2	0.07	0.03
N	Flow & PMMoV	Raw	0	0.08	0.04	O	Flow & PMMoV	Raw	0	0.08	0.04
N	Flow & PMMoV	Raw	1	0.10	0.05	O	Flow & PMMoV	Raw	1	0.09	0.04
N	Flow & PMMoV	Raw	2	0.14	0.14	O	Flow & PMMoV	Raw	2	0.14	0.13
N	Flow & PMMoV	Log	0	0.06	0.02	O	Flow & PMMoV	Log	0	0.06	0.02
N	Flow & PMMoV	Log	1	0.06	0.03	O	Flow & PMMoV	Log	1	0.06	0.03
N	Flow & PMMoV	Log	2	0.07	0.03	O	Flow & PMMoV	Log	2	0.07	0.03

Overall, the model with the lowest RMSE is the one using the virus concentrations of O, normalized by flow, with $\log_{10}$ transformation, and without any lags:

$$\log_{10}(C_{t+1}) = \beta_0 + \beta_1 \log_{10}(W_t^{\text{Flow}}) + \epsilon \tag{4}$$

where β_j's are the regression coefficients, and ϵ is the residual. I explored various approach to expand upon this model in the future sections.

3.2 Comparisons between Treatment Plants

Building a separate model for each WWTP is not very a parsimonious solution. It would be ideal if it is possible to develop one model using the pooled data from all WWTPs. To examine whether it is possible to do so, I compared the estimated error variance and the estimated regression coefficients of W_t^{Flow} of the model in (4) across WWTPs. If the estimated error variance and coefficients are similar, this would imply that the quality of the wastewater samples did not differ across WWTPs, and that the relationship between virus concentrations and case count did not differ across WWTPs. As a result, using one model for all WWTPs is justifiable. Otherwise, it may be more appropriate to use a separate model for each WWTP. Also, it would be interesting to examine whether a large estimated error

Table 3: Variance Estimators of the best model for each WWTP

Code	$\hat{\sigma}^2$	Code	$\hat{\sigma}^2$
AL	0.05	LF	0.06
AX	0.09	LS	0.10
BI	0.12	LY	0.68
BL	0.04	MH	0.30
BP	0.21	MK	0.02
CB	0.32	ML	0.07
CM	0.20	MR	0.10
CR	0.05	NB	0.10
EB	0.09	NF	0.04
EM	0.01	NP	0.04
EP	0.03	RC	0.02
ER	0.11	SD	0.02
FF	0.14	SN	0.02
GC	0.52	SV	0.02
GL	0.08	TC	0.01
HS	0.04	TR	0.14
HT	0.04	WA	0.04
HY	0.18	WL	0.05
IF	0.06	WM	0.05
LB	0.30	WT	0.06

variance was related to a lower estimated coefficient and a large p-value.

The variance estimator proposed by Rice (1984) was used in this analysis. For a given dataset $(x_1, y_1), ..., (x_n, y_n)$, where $x_1 \leq x_2 \leq \cdots \leq x_n$, the Rice's variance estimator is

$$\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (y_i - y_{i-1})^2 \tag{5}$$

Unlike residual-based variance estimators, Rice's variance estimator does not require an estimate of y. In this analysis, the x variable is $\log_{10}(W_t^{\text{Flow}})$ and the y variable is $\log_{10}(C_{t+1})$. The results are displayed in Table 3. The mean and standard deviation $\hat{\sigma}^2$ are 0.11 and 0.14 respectively. The $\hat{\sigma}^2$ is as low as 0.01 (EM and TC), and as high as 0.68 (LY).

Table 4 displays the estimated regression coefficients and p-values of $\log_{10}(W_t^{\text{Flow}})$ for each WWTP. It was found that the estimated coefficients ranged from -0.07 to 0.71 (mean = 0.28, SD = 0.23). Although the sign of the estimated coefficients in general did not differ across WWTPs, the magnitude varied a lot. The correlation between the variance estimate and the estimated regression coefficient was -0.59, and the correlation between the variance estimate and the p-value was 0.35. In other words, a larger variance estimate was related to a smaller estimated coefficient and a larger p-value.

Table 4: Estimated Coefficients and p-values of $\log_{10}(W_t^{\text{flow}}$ for each WWTP

Code	Estimated Coefficient	p-value	Code	Estimated Coefficient	p-value
AL	0.43	0.01	LF	0.17	0.38
AX	0.23	0.02	LS	0.30	0.08
BI	0.06	0.50	LY	-0.04	0.29
BL	0.51	0.00	MH	0.09	0.05
BP	0.22	0.35	MK	0.66	0.00
CB	0.02	0.73	ML	0.10	0.47
CM	0.11	0.34	MR	0.25	0.00
CR	0.28	0.13	NB	0.12	0.18
EB	0.18	0.06	NF	0.66	0.00
EM	0.62	0.00	NP	0.43	0.01
EP	0.58	0.00	RC	0.64	0.00
ER	0.00	0.94	SD	0.32	0.00
FF	0.13	0.26	SN	0.59	0.00
GC	0.03	0.65	SV	0.44	0.10
GL	0.17	0.34	TC	0.65	0.00
HS	0.24	0.07	TR	0.18	0.12
HT	0.22	0.24	WA	0.39	0.00
HY	0.04	0.82	WL	0.01	0.95
IF	0.00	0.98	WM	0.35	0.05
LB	-0.07	0.53	WT	0.71	0.00

3.3 Use of Vaccination Rate

In this section, I examined whether including vaccination rate is helpful with predictions. Let V_t^{one} be the percentage of people who have received at least one doses of vaccine out of the population served by a WWTP, and V_t^{full} be the percentage of fully vaccinated people. The models to be investigated were built upon the model in Equation (4). In addition to the normalized virus concentration W_t^{Flow}, the models also considered the main effects of V_t^{one} and/or V_t^{full}, as well as two- or three-way interactions between W_t^{Flow}, V_t^{one} and V_t^{full}. The models are listed in Table 5, along with their means and standard deviations of the cross-validated RMSE across WWTPs. It was found that the model including the two-way interaction between W_t^{Flow} and V_t^{one} had the most accurate predictions:

$$\log_{10}(C_{t+1}) = \beta_0 + \beta_1 \log_{10}(W_t^{\text{Flow}}) + \beta_2 \log_{10}(V_t^{\text{full}}) + \beta_3 \log_{10}(W_t^{\text{Flow}}) \cdot \log_{10}(V_t^{\text{full}}) + \epsilon \quad (6)$$

3.4 Use of Lagged Case Count

In this section, I examined whether virus concentrations were still a useful predictor for predicting C_{t+1}, if the lagged case count C_t was included in the model. I compared the pre-

Table 5: Means and standard deviations of RMSE across WWTPs in all 48 conditions

Formula	Mean	SD
$\log_{10}(C_{t+1}) \sim \log_{10}(W_t^{\text{flow}}) + \log_{10}(V_t^{\text{one}})$	0.06	0.02
$\log_{10}(C_{t+1}) \sim \log_{10}(W_t^{\text{flow}}) + \log_{10}(V_t^{\text{full}})$	0.05	0.02
$\log_{10}(C_{t+1}) \sim \log_{10}(W_t^{\text{flow}}) + \log_{10}(V_t^{\text{one}}) + \log_{10}(V_t^{\text{full}})$	0.05	0.02
$\log_{10}(C_{t+1}) \sim \log_{10}(W_t^{\text{flow}}) * \log_{10}(V_t^{\text{one}})$	0.06	0.02
$\log_{10}(C_{t+1}) \sim \log_{10}(W_t^{\text{flow}}) * \log_{10}(V_t^{\text{full}})$	0.05	0.02
$\log_{10}(C_{t+1}) \sim \log_{10}(W_t^{\text{flow}}) * \log_{10}(V_t^{\text{one}}) * \log_{10}(V_t^{\text{full}})$	0.09	0.08

Table 6: Means and standard deviations of RMSE across WWTPs when using virus concentrations only, lagged case count only, and both in the model

Model	Mean	SD
$\log_{10}(C_{t+1}) \sim \log_{10}(C_t)$	0.05	0.02
$\log_{10}(C_{t+1}) \sim \log_{10}(W_t^{\text{flow}}) * \log_{10}(V_t^{\text{full}})$	0.05	0.02
$\log_{10}(C_{t+1}) \sim \log_{10}(W_t^{\text{flow}}) * \log_{10}(V_t^{\text{full}}) + \log_{10}(C_t)$	0.05	0.02

dictive performance of a model that only used C_t, a model that used the two-way interaction between W_t^{Flow} and V_t^{one}, and a model that used both.

The models were again fitted separately for each WWTP and evaluated by a LOOCV. The results are presented in Table 6. The model using the interaction between W_t^{Flow} and V_t^{one} had the lowest cross-validated RMSE, but the model that only used C_t was able to achieve a very similar RMSE.

3.5 Prediction Accuracy Over Different Forecast Horizons

In this section, I assessed the performance of predictions over different forecast horizons (same week, one week ahead and two weeks ahead). The same model in Equation (6) was fitted separately for each WWTP, with $\log_{10}(C_t), \log_{10}(C_{t+1})$ and $\log_{10}(C_{t+2})$ being the dependent variable. The means and standard deviations of the cross-validated RMSE are reported in the first three rows of Table 7. Interestingly, the results demonstrate that predicting case count one week ahead is more accurate than predicting case count of the same week. The predictions for two weeks ahead is substantially worse than the predictions for the first week. For comparisons, I fitted the model in Equation (4) again with different forecast horizons. The results are more in line with what one would expect: the longer the forecast horizon, the more accurate the predictions (last three rows of Table 7).

Table 7: Means and standard deviations of RMSE across WWTPs when predicting the case count of the same week, one week ahead and two weeks ahead

Model	Mean	SD
Using W_t^{flow} and V_t^{full}		
$\log(C_t) \sim \log_{10}(W_t^{\text{flow}}) * \log_{10}(V_t^{\text{full}})$	0.06	0.02
$\log(C_{t+1}) \sim \log_{10}(W_t^{\text{flow}}) * \log_{10}(V_t^{\text{full}})$	0.05	0.02
$\log(C_{t+2}) \sim \log_{10}(W_t^{\text{flow}}) * \log_{10}(V_t^{\text{full}})$	0.09	0.19
Using W_t^{flow} only		
$\log(C_t) \sim \log(W_t^{\text{flow}})$	0.06	0.02
$\log(C_{t+1}) \sim \log(W_t^{\text{flow}})$	0.06	0.02
$\log(C_{t+2}) \sim \log(W_t^{\text{flow}})$	0.07	0.02

4 Conclusion

Consistent with some studies in the wastewater literature (Maal-Bared et al., 2023; Feng et al., 2021; Duvallet et al., 2022), the current study found that the predictive performance did not improve after the normalization by either PMMoV alone or both PMMoV and flow. In fact, the prediction accuracy decreased, unless a $\log_{10}$ transformation was used. Normalization by flow was insensitive to the scale of the data. Moreover, the current findings suggest that the concentrations of either N or O can be used, as they resulted in similar predictions, most likely because they are highly correlated. Adding lagged virus concentrations to the model generally worsened the predictive performance. Using $\log_{10}$ transformation tended to improve the predictive performance.

The estimated error variance and regression coefficient varied substantially across WWTPs. These findings suggest that the relationship between COVID-19 incidence and SARS-CoV-2 RNA in wastewater may be treatment plant-specific, and future work will need to continue investigating how to appropriately normalize data from different plants to allow for cross-plant comparisons. Additionally, this suggests that at present, COVID-19 WBE may need to be validated at individual plants.

Using the vaccination rate in the model has been shown to improve the predictions. However, the results are not very consistent across different forecast horizons. It is possible that it is because the data are all collected in 2022, where the vaccination rate has already slowed down. People who want to be vaccinated have already been vaccinated, and people who don't are unlikely to change their minds. The restricted range of this variable may result in overfitting and an underestimation of its relationship with the case count.

A somewhat discouraging result is that using lagged case count alone provided similar predictive performance to using both virus concentrations and vaccination rates. More research should be done to evaluate whether WBE is actually worthwhile.

One shortcoming of this study is that the sampling period is rather short, and hence the available data are rather limited. The number of samples from each WWTP was mostly around 30. Moreover, the pandemic had been already slowing down during the sampling period. No data were available during any of the epidemic waves or lockdown periods in 2020 and 2021. Therefore, it is unclear that whether the current findings can be generalized to these situations.

Reference

Ahmed, W., Angel, N., Edson, J., Bibby, K., Bivins, A., O'Brien, J. W., Choi, P. M., Kitajima, M., Simpson, S. L., Li, J., et al. (2020). First confirmed detection of sars-cov-2 in untreated wastewater in australia: a proof of concept for the wastewater surveillance of covid-19 in the community. *Science of the Total Environment*, 728:138764.

Arora, S., Nag, A., Sethi, J., Rajvanshi, J., Saxena, S., Shrivastava, S. K., and Gupta, A. B. (2020). Sewage surveillance for the presence of sars-cov-2 genome as a useful wastewater based epidemiology (wbe) tracking tool in india. *Water Science and Technology*, 82(12):2823–2836.

Barasa, E. W., Ouma, P. O., and Okiro, E. A. (2020). Assessing the hospital surge capacity of the kenyan health system in the face of the covid-19 pandemic. *PLoS One*, 15(7):e0236308.

Chan, V. W.-S., Chiu, P. K.-F., Yee, C.-H., Yuan, Y., Ng, C.-F., and Teoh, J. Y.-C. (2021). A systematic review on covid-19: urological manifestations, viral rna detection and special considerations in urological conditions. *World journal of urology*, 39(9):3127–3138.

Chen, Y., Chen, L., Deng, Q., Zhang, G., Wu, K., Ni, L., Yang, Y., Liu, B., Wang, W., Wei, C., et al. (2020). The presence of sars-cov-2 rna in the feces of covid-19 patients. *Journal of medical virology*, 92(7):833–840.

Cheung, K. S., Hung, I. F., Chan, P. P., Lung, K., Tso, E., Liu, R., Ng, Y., Chu, M. Y., Chung, T. W., Tam, A. R., et al. (2020). Gastrointestinal manifestations of sars-cov-2 infection and virus load in fecal samples from a hong kong cohort: systematic review and meta-analysis. *Gastroenterology*, 159(1):81–95.

Duvallet, C., Wu, F., McElroy, K. A., Imakaev, M., Endo, N., Xiao, A., Zhang, J., Floyd-O'Sullivan, R., Powell, M. M., Mendola, S., et al. (2022). Nationwide trends in covid-19 cases and sars-cov-2 rna wastewater concentrations in the united states. *ACS ES&T Water*.

Farkas, K., Pellett, C., Alex-Sanders, N., Bridgman, M. T., Corbishley, A., Grimsley, J. M., Kasprzyk-Hordern, B., Kevill, J. L., Pântea, I., Richardson-O'Neill, I. S., et al. (2022). Comparative assessment of filtration-and precipitation-based methods for the concentration of sars-cov-2 and other viruses from wastewater. *Microbiology spectrum*, 10(4):e01102–22.

Feng, S., Roguet, A., McClary-Gutierrez, J. S., Newton, R. J., Kloczko, N., Meiman, J. G., and McLellan, S. L. (2021). Evaluation of sampling, analysis, and normalization methods for sars-cov-2 concentrations in wastewater to assess covid-19 burdens in wisconsin communities. *Acs Es&T Water*, 1(8):1955–1965.

Hamza, H., Rizk, N. M., Gad, M. A., and Hamza, I. A. (2019). Pepper mild mottle virus in wastewater in egypt: a potential indicator of wastewater pollution and the efficiency of the treatment process. *Archives of Virology*, 164(11):2707–2713.

Hasan, S. W., Ibrahim, Y., Daou, M., Kannout, H., Jan, N., Lopes, A., Alsafar, H., and Yousef, A. F. (2021). Detection and quantification of sars-cov-2 rna in wastewater and treated effluents: Surveillance of covid-19 epidemic in the united arab emirates. *Science of The Total Environment*, 764:142929.

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

Kitajima, M., Iker, B. C., Pepper, I. L., and Gerba, C. P. (2014). Relative abundance and treatment reduction of viruses during wastewater treatment processes—identification of potential viral indicators. *Science of the Total Environment*, 488:290–296.

Kitajima, M., Sassi, H. P., and Torrey, J. R. (2018). Pepper mild mottle virus as a water quality indicator. *NPJ Clean Water*, 1(1):1–9.

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., and Lessler, J. (2020). The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 172(9):577–582.

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., and Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov-2). *Science*, 368(6490):489–493.

Maal-Bared, R., Qiu, Y., Li, Q., Gao, T., Hrudey, S. E., Bhavanam, S., Ruecker, N. J., Ellehoj, E., Lee, B. E., and Pang, X. (2023). Does normalization of sars-cov-2 concentrations by pepper mild mottle virus improve correlations and lead time between wastewater surveillance and clinical data in alberta (canada): comparing twelve sars-cov-2 normalization approaches. *Science of The Total Environment*, 856:158964.

Manor, Y., Handsher, R., Halmut, T., Neuman, M., Bobrov, A., Rudich, H., Vonsover, A., Shulman, L., Kew, O., and Mendelson, E. (1999). Detection of poliovirus circulation by environmental surveillance in the absence of clinical cases in israel and the palestinian authority. *Journal of clinical microbiology*, 37(6):1670–1675.

Nishiura, H., Kobayashi, T., Miyama, T., Suzuki, A., Jung, S.-m., Hayashi, K., Kinoshita, R., Yang, Y., Yuan, B., Akhmetzhanov, A. R., et al. (2020). Estimation of the asymptomatic ratio of novel coronavirus infections (covid-19). *International journal of infectious diseases*, 94:154–155.

Oran, D. P. and Topol, E. J. (2020). Prevalence of asymptomatic sars-cov-2 infection: a narrative review. *Annals of internal medicine*, 173(5):362–367.

Parasa, S., Desai, M., Chandrasekar, V. T., Patel, H. K., Kennedy, K. F., Roesch, T., Spadaccini, M., Colombo, M., Gabbiadini, R., Artifon, E. L., 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–e2011335.

Peccia, J., Zulli, A., Brackney, D. E., Grubaugh, N. D., Kaplan, E. H., Casanovas-Massana, A., Ko, A. I., Malik, A. A., Wang, D., Wang, M., et al. (2020). Measurement of sars-cov-2 rna in wastewater tracks community infection dynamics. *Nature biotechnology*, 38(10):1164–1167.

Post, L. A., Issa, T. Z., Boctor, M. J., Moss, C. B., Murphy, R. L., Ison, M. G., Achenbach, C. J., Resnick, D., Singh, L. N., White, J., et al. (2020). Dynamic public health surveillance to track and mitigate the us covid-19 epidemic: longitudinal trend analysis study. *Journal of medical Internet research*, 22(12):e24286.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Randazzo, W., Truchado, P., Cuevas-Ferrando, E., Simón, P., Allende, A., and Sánchez, G. (2020). Sars-cov-2 rna in wastewater anticipated covid-19 occurrence in a low prevalence area. *Water research*, 181:115942.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics*, pages 1215–1230.

Rosario, K., Symonds, E. M., Sinigalliano, C., Stewart, J., and Breitbart, M. (2009). Pepper mild mottle virus as an indicator of fecal pollution. *Applied and environmental microbiology*, 75(22):7261–7267.

Weidhaas, J., Aanderud, Z. T., Roper, D. K., VanDerslice, J., Gaddis, E. B., Ostermiller, J., Hoffman, K., Jamal, R., Heck, P., Zhang, Y., et al. (2021). Correlation of sars-cov-2 rna in wastewater with covid-19 disease burden in sewersheds. *Science of The Total Environment*, 775:145790.

Wong, M. C., Huang, J., Lai, C., Ng, R., Chan, F. K., and Chan, P. K. (2020). Detection of sars-cov-2 rna in fecal specimens of patients with confirmed covid-19: a meta-analysis. *Journal of Infection*, 81(2):e31–e38.

Ye, Y., Ellenberg, R. M., Graham, K. E., and Wigginton, K. R. (2016). Survivability, partitioning, and recovery of enveloped viruses in untreated municipal wastewater. *Environmental science & technology*, 50(10):5077–5085.

Zhan, Q., Babler, K. M., Sharkey, M. E., Amirali, A., Beaver, C. C., Boone, M. M., Comerford, S., Cooper, D., Cortizas, E. M., Currall, B. B., et al. (2022). Relationships between sars-cov-2 in wastewater and covid-19 clinical cases and hospitalizations, with and without normalization against indicators of human waste. *ACS ES&T Water*.

Zhang, T., Breitbart, M., Lee, W. H., Run, J.-Q., Wei, C. L., Soh, S. W. L., Hibberd, M. L., Liu, E. T., Rohwer, F., and Ruan, Y. (2006). Rna viral community in human feces: prevalence of plant pathogenic viruses. *PLoS biology*, 4(1):e3.