Advanced Regression Techniques for Prediction of Covid-19 Case Counts in Minnesota Using RNA Data

Kah Meng Soh
University of Minnesota - Twin Cities

Abstract

This study builds on the initial research, 'Wastewater Surveillance of SARS-CoV-2 in Minnesota (Mark et al. 2024. Multidisciplinary Digital Publishing Institute),' which applied various linear regression models to individual wastewater treatment plants. By aggregating data from all plants across Minnesota, this research aims to improve the predictive accuracy of COVID-19 case counts. We employed advanced statistical methods, including interaction effects, regularization, and Generalized Additive Models (GAMs), along with machine learning techniques such as Decision Trees, K-Nearest Neighbors, and Gradient Boosting Machines (GBM). Model diagnostics also performed on the best linear model to check if linear model assumption are met. The study evaluates the effectiveness of normalization versus interaction effects and examines the impact of using lagged data (lag 1 and lag 2) on model performance. The results indicate that K-Nearest Neighbors (KNN) with k=9 outperforms other models, achieving the lowest prediction error. Additionally, the research finds that excluding PMMoV improves model performance, interaction effects are preferable to normalization, and using only the most recent data (lag 1) is sufficient compared to incorporating both lag 1 and lag 2.

Background

Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has infected millions of people worldwide, resulting in significant health and economic impacts. Consequently, methods for detecting, tracking, and sampling 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, Ouma, and Okiro 2020). Additionally, not everyone has access to testing due to economic, geographic, or social restrictions. Test results are also a lagging indicator of the pandemic's progression because testing is usually prompted by symptoms, which may take up to two weeks to appear 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). 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 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). Regarding 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 is shed in large quantities in wastewater (Hamza et al. 2019; Kitajima et al. 2014; Kitajima, Sassi, and Torrey 2018; Rosario et al. 2009; Zhang et al. 2006). It is also highly stable in wastewater, and its concentrations show little seasonal variation (Kitajima et al. 2014; Kitajima, Sassi, and Torrey 2018).

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

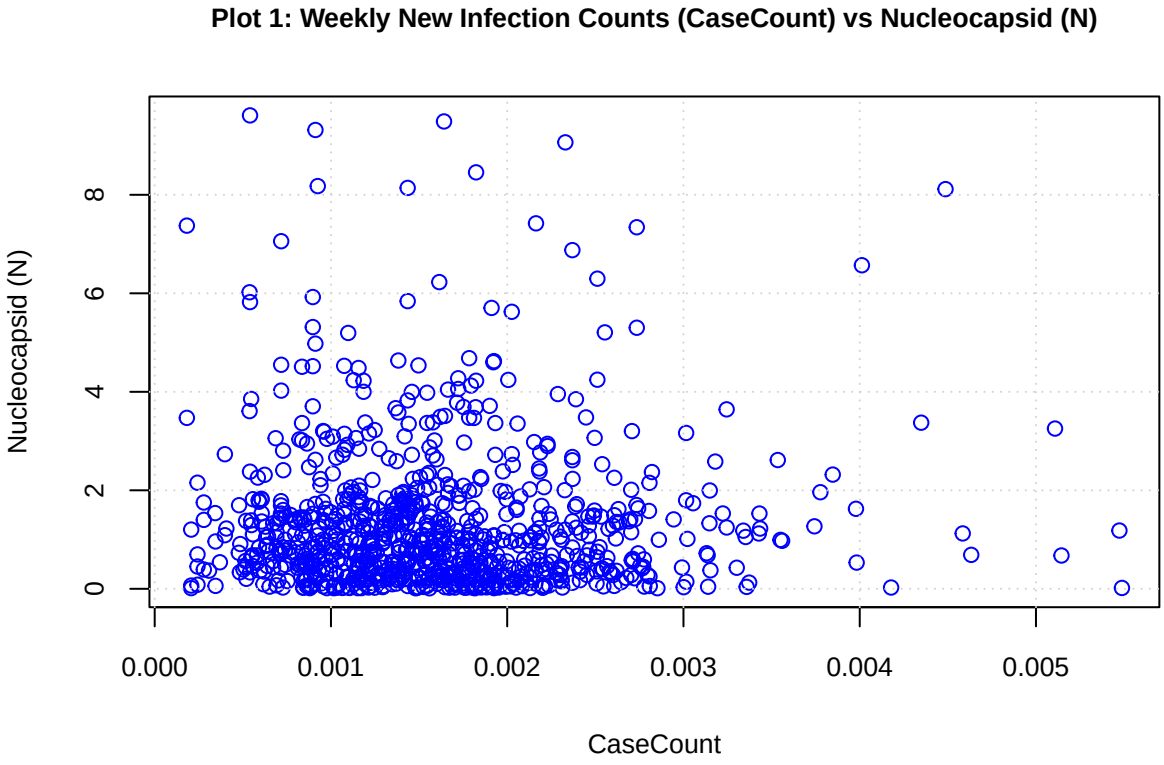
1. Should the SARS-CoV-2 concentrations be normalized using Flow, PMMoV, or the interaction between them?
2. Is linear model assumption satisfied?
3. How do linear model compare to machine learning models in predicting outcomes?
4. Should the model incorporate only the lagged 1 data, or should both lagged 1 and lagged 2 data be included?

Data Description

Wastewater samples were collected from 40 wastewater treatment plants (WWTPs) across 191 zip codes in Minnesota, covering approximately 67% of the state's population, from March 2022 to October 2022. Samples were taken twice weekly, and the concentrations of SARS-CoV-2 target genes, specifically Nucleocapsid (N), were measured as gene copies per liter. Weekly averages were used for data since two samples were collected weekly. Data inclusion was limited to non-overlapping weeks, and missing values were excluded. Each weekly dataset also included measurements of the human fecal marker (PMMoV) and influent flow rate (Flow) provided by the participating WWTPs.

Weekly new infection counts (CaseCount) for each WWTP service area were obtained from Minnesota's public health records. To ensure consistent data scaling, CaseCount and Nucleocapsid (N) were divided by the population size of each respective WWTP service area, resulting in CaseCount being expressed as the percentage of the population infected with COVID-19. Observations with missing values due to the creation of lagged variables were removed. The final dataset comprises 930 weeekly samples data. Previous studies determined that applying a $\log_{10}$ transformation to both the dependent and independent variables enhances model performance, so this transformation will be maintained in our analysis.

Plot 1 is a scatterplot displaying Weekly New Infection Counts (CaseCount) versus Nucleocapsid (N). To improve visibility, extreme outlier values have been removed. The data is shown on the original scale, without a log transformation, as the log transformation is monotonic in nature to the original scale. The plot reveals no obvious linear trend, as many data points cluster around lower values of CaseCount and Nucleocapsid. This suggests that a nonlinear model might better fit the data. We will perform linear model diagnostics and explore nonlinear modeling to further investigate this possibility.



Statistical Modelling

Denote C_t as the COVID-19 case count rate at time t, N_t as the SARS-CoV-2 concentrations at time t, $PMMoV_t$ as the human fecal marker concentration at time t, and $Flow_t$ as the influent flow rate at time t. Following the methodology of the original study, we will use Root Mean Square Error (RMSE) to evaluate the performance of each model because it aligns with the scale of the original target, the COVID-19 case count rate. However, instead of using Leave-One-Out Cross-Validation (LOOCV), we will implement 10-fold cross-validation

due to its greater time efficiency and suitability for typical statistical studies. To do the cross-validation on the original scale, we will reverse transform via exponentiating to ensure predictions are made based on the original data scale, thus avoiding data leakage. Unlike the original study, which modeled each individual plant separately, this study treated all the data collected in Minnesota as one aggregated dataset. This approach allows for predictions at the population level rather than at the level of individual WWTPs.

Considering the bias-variance trade-off inherent in model selection:

a) Bias: Bias refers to errors introduced by assuming that the model is too simple. A high-bias model might not capture all the complexities of the data, leading to underfitting.
b) Variance: Variance refers to errors introduced by the model being too complex. A high-variance model might fit the training data too closely, capturing noise as if it were a real pattern, leading to overfitting.

We will explore various statistical and machine learning methods to determine if a complex non-linear relationship between case counts and virus concentrations provides a better fit than a simple linear model.

All statistical analyses were performed using R version 4.4.1 (R Core Team, 2024).

Linear Regression

Linear regression is a fundamental statistical model that assumes a linear relationship between the predictors and the response. This model is typically characterized by high bias and low variance due to its linearity assumption. Table 1 presents the RMSE for various linear regression models, including those using only lagged 1 predictors and those using both lagged 1 and lagged 2 predictors, across different normalization scenarios: unnormalized, normalized by flow, and normalized by both flow and PMMoV.

From **Table 1**, based on the RMSE values, we observed similar performance between models using lagged 1 predictors and those using both lagged 1 and lagged 2 predictors across all terms. Our analysis indicates that the best model fit is achieved by normalizing only by flow, followed by normalizing by both flow and PMMoV. Normalizing by only PMMoV is next, with the unnormalized model showing the least fit.

Since the inclusion of PMMoV does not improve the model, aligning with findings from the original paper, we conclude that PMMoV should be excluded from future model considerations. Henceforth, all future models will exclude PMMoV.

Table 1: Linear Regression Model Performance Using RMSE

Model	RMSE
$\log_{10}(C_t) \sim \log_{10}(N_{t-1})$	0.0007808
$\log_{10}(C_t) \sim \log_{10}(N_{t-1}) + \log_{10}(N_{t-2})$	0.0007779
$\log_{10}(C_t) \sim \log_{10}(N_{t-1} \cdot Flow_{t-1})$	0.0007350
$\log_{10}(C_t) \sim \log_{10}(N_{t-1} \cdot Flow_{t-1}) + \log_{10}(N_{t-2} \cdot Flow_{t-2})$	0.0007331
$\log_{10}(C_t) \sim \log_{10}(\frac{N_{t-1}}{PMMoV_{t-1}})$	0.0007628
$\log_{10}(C_t) \sim \log_{10}(\frac{N_{t-1}}{PMMoV_{t-1}}) + \log_{10}(\frac{N_{t-2}}{PMMoV_{t-2}})$	0.0007623
$\log_{10}(C_t) \sim \log_{10}(\frac{N_{t-1} \cdot Flow_{t-1}}{PMMoV_{t-1}})$	0.0007368
$\log_{10}(C_t) \sim \log_{10}(\frac{N_{t-1} \cdot Flow_{t-1}}{PMMoV_{t-1}}) + \log_{10}(\frac{N_{t-2} \cdot Flow_{t-2}}{PMMoV_{t-2}})$	0.0007350

Table 2: Linear Regression with Interaction Model Performance Using RMSE

Model	RMSE
$\log_{10}(C_t) \sim \log_{10}(N_{t-1}) * \log_{10}(Flow_{t-1})$	0.0007298
$\log_{10}(C_t) \sim \log_{10}(N_{t-1}) * \log_{10}(Flow_{t-1}) + \log_{10}(N_{t-2}) * \log_{10}(Flow_{t-2})$	0.0007274

Linear Regression with Interaction Effect

For the best models from Table 1 with the lowest RMSE, we now consider using the interaction term between Flow and SARS-CoV-2 (N) instead of normalizing. In statistical terms, an interaction term means that the effect of one predictor on the response depends on another predictor. Table 2 presents the RMSE for the three models using interaction terms instead of normalization.

From **Table 2**, based on the RMSE values, we observed that models with interaction terms performed better than those with normalization, for both lagged 1 predictors and models using both lagged 1 and lagged 2 predictors. Therefore, we conclude that incorporating interaction terms is superior to normalization. Although models with both lagged 1 and lagged 2 predictors showed slightly better performance than those with only lagged 1 predictors, the increased model complexity prevents us from definitively determining whether using both lagged 1 and lagged 2 predictors is better. For educational purposes, we will continue to model with normalization in all future models.

Table 3: Linear Regression Model Performance with RMSE using Ridge Regression

Model	RMSE
$\log_{10}(C_t) \sim \log_{10}(N_{t-1} \cdot Flow_{t-1}) + \log_{10}(N_{t-2} \cdot Flow_{t-2})$	0.0007349
$\log_{10}(C_t) \sim \log_{10}(N_{t-1}) * \log_{10}(Flow_{t-1}) + \log_{10}(N_{t-2}) * \log_{10}(Flow_{t-2})$	0.0007305

Regularized Linear Regression Using Ridge Method

Now, we will try regularized linear regression using Ridge. Ridge is a method that can handle collinearity between predictors by shrinking all correlated predictors' coefficients together. This approach makes sense for our data, as our predictors are very similar to each other (e.g., lagged 1 and lagged 2). Although the correlation between N1 lagged 1 and N1 lagged 2 is only 0.25, we are still interested in understanding how regularization would work.

We use the glmnet package, which begins by specifying a range of lambda values. Next, through 10-fold cross-validation, the algorithm identifies the optimal lambda for each fold, resulting in 10 distinct lambda values. These optimal lambdas are then evaluated on the entire dataset again to select the lambda that minimizes the prediction error. Because fitting models with Ridge for only one predictor is equivalent to fitting ordinary linear regression, we will fit models with both lagged 1 and lagged 2 terms.

From **Table 3**, we see that Ridge performs worse than the linear model with interaction term and hence should not be considered further.

Model Diagostic for Linear Model Assumption

Plot 2 showcases the residuals vs. fitted plot and normal Q-Q plot for the model with interaction effects, considering both lagged 1 and lagged 2 predictors, note that this is model 2 in **Table 2** without ridge regularized.

$$\log_{10}(C_t) \sim \log_{10}(N_{t-1}) * \log_{10}(\text{Flow}_{t-1}) + \log_{10}(N_{t-2}) * \log_{10}(\text{Flow}_{t-2})$$

Residuals vs. Fitted Values Plot

Linearity: This plot helps check if the relationship between the predictors and the response is linear. Ideally, the residuals should be randomly scattered around the horizontal line (y = 0) without any discernible pattern. From the plot, it appears there is no clear pattern, indicating that the linearity assumption is reasonably met.

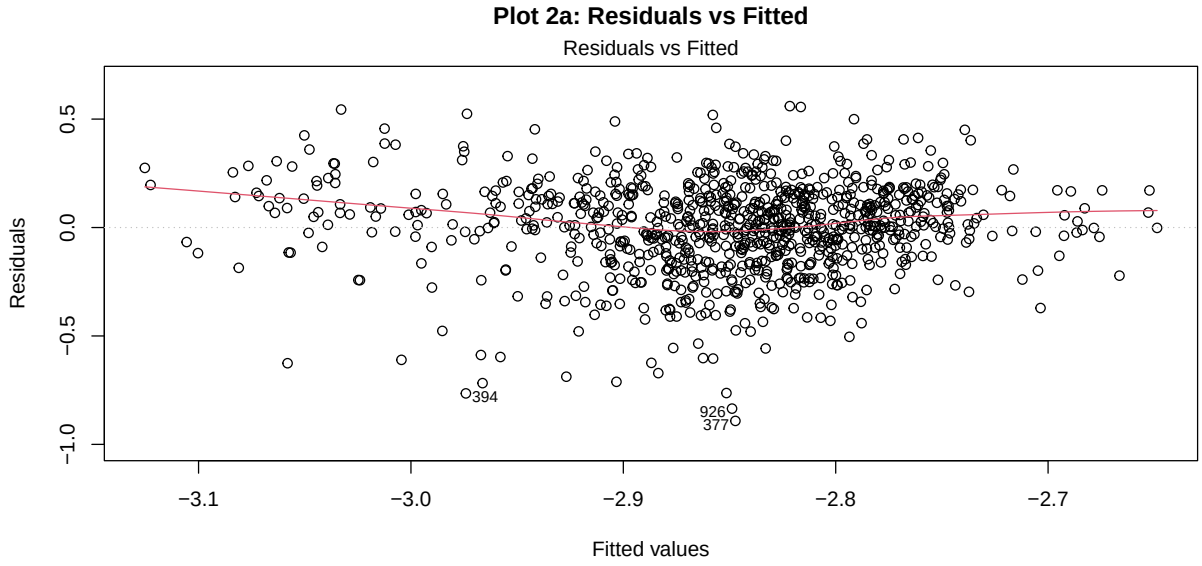
Homoscedasticity: This refers to the constant variance of the residuals. The residuals should form a horizontal band around the centerline (y = 0) with constant spread. The spread of the residuals seems fairly constant, though there might be slight funneling on the right side. This suggests that homoscedasticity is mostly met, but there may be some minor issues.

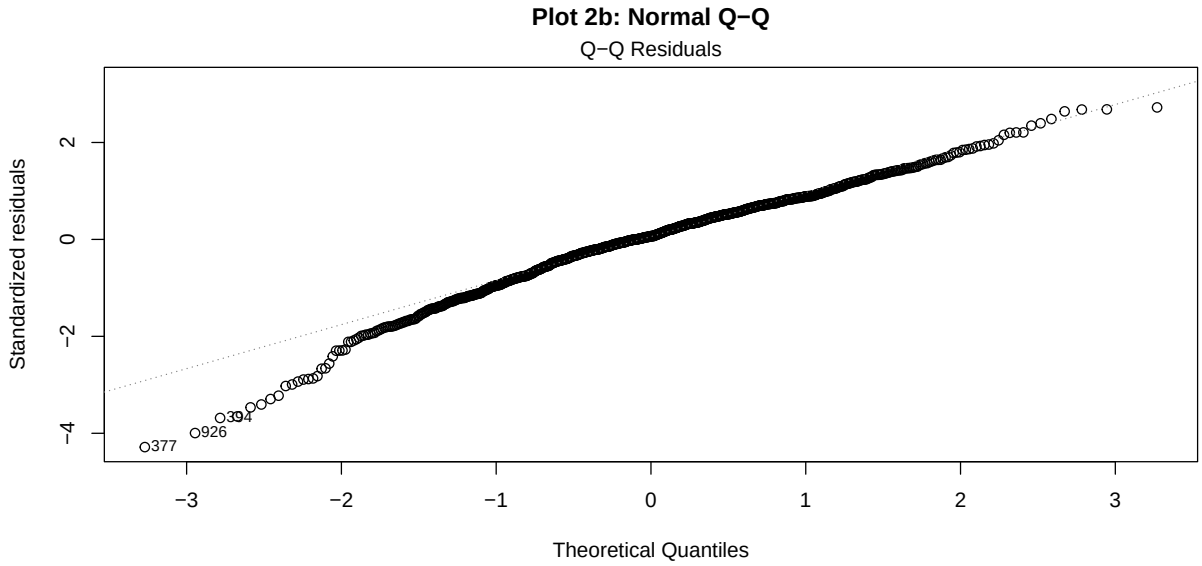
Q-Q Plot of Residuals

Normality of Residuals: The Q-Q plot assesses if the residuals follow a normal distribution. The points should lie along the 45-degree reference line. The Q-Q plot shows that most points lie along the reference line, indicating that the residuals are approximately normally distributed. There are some deviations at the tails, which are common in real data but should be checked for potential outliers or heavy-tailed distribution.

- Linearity: Assumption appears to be met.
- Homoscedasticity: Mostly met with some minor issues.
- Normality: Mostly met with some deviations at the tails.

Overall, the linear model assumptions are reasonably satisfied. Moving forward, we will consider some nonlinear modeling techniques to compare their performance with the linear model.





Generalized Additive Model

We now considered some nonlinear modeling with Generalized Additive Models (GAMs). Although GAMs are still parametric, they allow us to fit smooth functions between predictors and the response, effectively modeling nonlinearity. We use the mgcv package, which automates the process of finding the best model fit in each of the 10 folds. This includes automatically selecting or optimizing smoothing parameters (such as degrees of freedom or lambda values) using techniques like Generalized Cross Validation (GCV) or Restricted Max-

Table 4: GAM Model Performance Using RMSE

Model	RMSE
$\log_{10}(C_t) \sim s(\log_{10}(N_{t-1}), by = \log_{10}(Flow_{t-1}))$	0.0007450
$\log_{10}(C_t) \sim s(\log_{10}(N_{t-1}), by = \log_{10}(Flow_{t-1})) + s(\log_{10}(N_{t-2}), by = \log_{10}(Flow_{t-2}))$	0.0007423
$\log_{10}(C_t) \sim s(\log_{10}(N_{t-1} \cdot Flow_{t-1}))$	0.0007117
$\log_{10}(C_t) \sim s(\log_{10}(N_{t-1} \cdot Flow_{t-1})) + s(\log_{10}(N_{t-2} \cdot Flow_{t-2}))$	0.0007093

imum Likelihood (REML). The package computes errors and averages these errors across folds to provide a final evaluation.

From **Table 4**, we observe that GAMs with interaction terms perform the best so far, indicating that nonlinear modeling may fit the data better. Interestingly, interaction terms work better than normalization for GAMs, which is surprising because the complexity of the GAM model is often not well-suited for capturing interactions effectively. While it is possible to write out the equation for GAMs since they are parametric models, we have omitted it here due to its complexity and because later models provide better performance.

Decision Tree Regression

Now, we will move on to machine learning models that cover nonlinear modeling. We'll start with basic tree-based methods, specifically the decision tree. Although decision trees are commonly used for classification, they can also be applied to regression, where predictions are based on the mean of the data in the split regions.

There are several hyperparameters for tuning a decision tree, such as:

- Maximum Depth: The maximum depth of the tree.
- Minimum Samples Split: The minimum number of samples required to split an internal node.
- Minimum Samples Leaf: The minimum number of samples required to be at a leaf node.

From **Table 5**, we see that the decision tree performs worse than the previous models and, therefore, should not be considered further. Therefore we doesn't bother with the hyperparameter selected by caret package with optimized the tree for best result.

Table 5: Decision Tree Regression Model Performance Using RMSE

Model	RMSE
$\log_{10}(C_t) \sim \log_{10}(N_{t-1} \cdot Flow_{t-1})$	0.0007224
$\log_{10}(C_t) \sim \log_{10}(N_{t-1} \cdot Flow_{t-1}) + \log_{10}(N_{t-2} \cdot Flow_{t-2})$	0.0007224
$\log_{10}(C_t) \sim \log_{10}(N_{t-1}) * \log_{10}(Flow_{t-1})$	0.0007466
$\log_{10}(C_t) \sim \log_{10}(N_{t-1}) * \log_{10}(Flow_{t-1}) + \log_{10}(N_{t-2}) * \log_{10}(Flow_{t-2})$	0.0007396

Gradient Boosting Regression

Gradient Boosting Regression is a boosting and ensemble method, which means training many weak models that learn from previous mistakes to form a strong model. In our case, we will use decision trees as the weak models.

The following are explanations of the hyperparameters for Gradient Boosting according to the caret package:

- n.trees (or n_estimators): This parameter determines the number of trees to be built in the model. Each tree is built sequentially to correct the errors of the previous trees.
- interaction.depth (or max_depth): This parameter controls the maximum depth of each tree. It limits the number of splits in each tree, thus controlling the model's complexity.
- shrinkage (or learning_rate): This parameter controls the contribution of each tree to the final model. A lower learning rate means the model learns slowly but can result in better generalization by making finer adjustments.
- n.minobsinnode (or min_samples_split / min_child_weight): This parameter sets the minimum number of observations that must be present in a node for it to be split. It prevents the model from learning overly specific patterns (overfitting) by ensuring that each split is based on enough data.

It's important to note how the package 'caret' fine-tunes hyperparameters for both GBMs and KNN. Let's take KNN as an example:

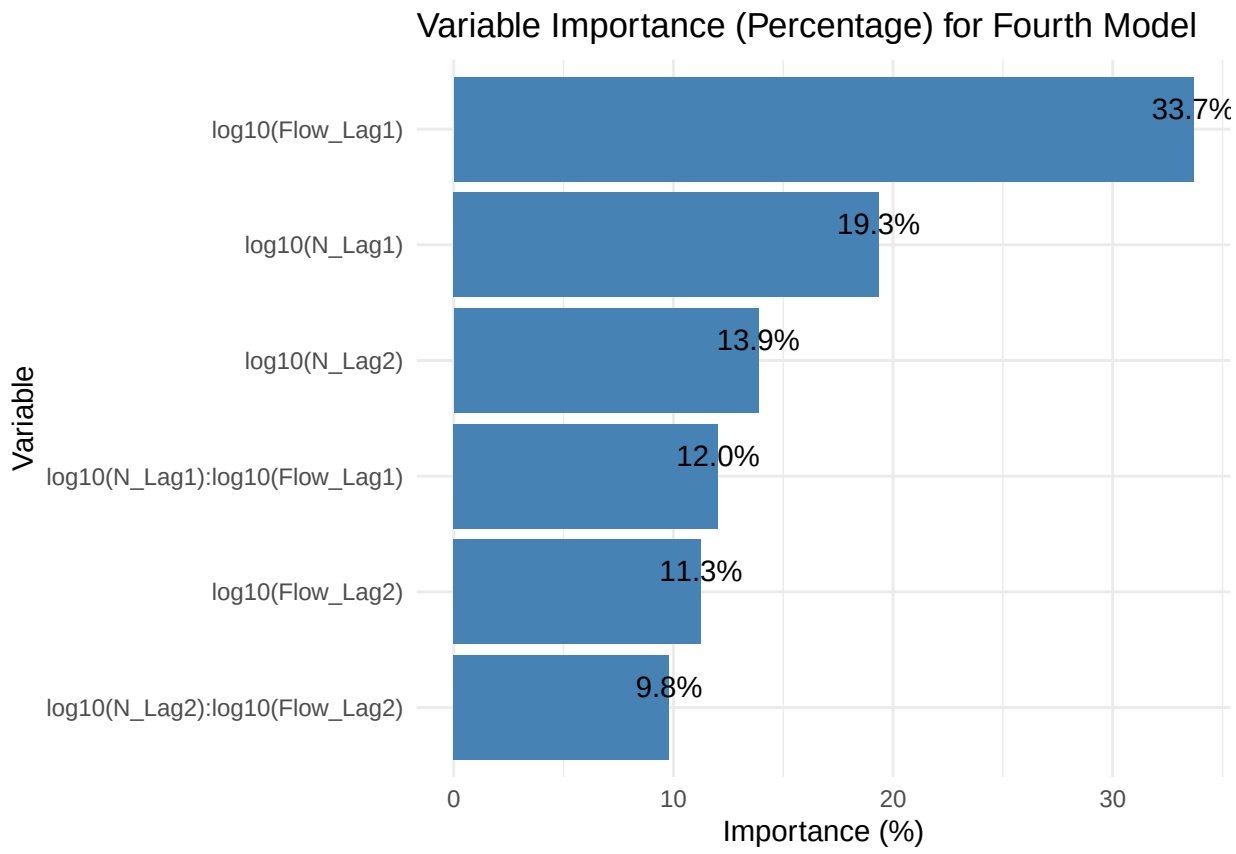
The data is divided into 10 folds, various k values are tried for each fold, performance metrics (e.g., RMSE) are calculated and averaged across all folds for each k value, and the k with the lowest average RMSE is selected to train the final model on the entire dataset. For GBMs, since it has multiple hyperparameters, a combination of hyperparameter values is used.

From **Table 6**, we observe that Gradient Boosting Regression surpassed the previous model when using interaction terms instead of normalization. This model preferred incorporating

Table 6: GBM Model Performance Using RMSE

Model	RMSE	BestHyperparameters
$\log_{10}(C_t) \sim \log_{10}(N_{t-1} \cdot Flow_{t-1})$	0.0007033	n.trees=50, interaction.depth=1, shrinkage=0.1, n.minobsinnode=10
$\log_{10}(C_t) \sim \log_{10}(N_{t-1} \cdot Flow_{t-1}) + \log_{10}(N_{t-2} \cdot Flow_{t-2})$	0.0007007	n.trees=50, interaction.depth=1, shrinkage=0.1, n.minobsinnode=10
$\log_{10}(C_t) \sim \log_{10}(N_{t-1}) * \log_{10}(Flow_{t-1})$	0.0006999	n.trees=100, interaction.depth=1, shrinkage=0.1, n.minobsinnode=10
$\log_{10}(C_t) \sim \log_{10}(N_{t-1}) * \log_{10}(Flow_{t-1}) + \log_{10}(N_{t-2}) * \log_{10}(Flow_{t-2})$	0.0006654	n.trees=100, interaction.depth=2, shrinkage=0.1, n.minobsinnode=10

both lagged 1 and lagged 2 predictors, as demonstrated by the variable importance plot. The plot assigns higher importance to lagged 1 predictors compared to lagged 2 predictors, indicating that lagged 1 predictors significantly reduce the average error more than lagged 2 predictors.



K-Nearest Neighbours Regression

For the final machine learning model, we employed K-Nearest Neighbors (KNN). KNN sets decision boundaries based on the majority of nearby data points. In K-Nearest Neighbors Regression, predictions are made by averaging the values of the nearby data points. The primary advantage of KNN is that it has only one hyperparameter, k, which represents the number of nearby data points considered.

From Table 7, we see that K-Nearest Neighbors performed the best of all models when using the interaction term over normalization terms. The results between using only lagged 1 and both lagged 1 and lagged 2 predictors are very close. With supporting evidence from previous variable importance plots from GBMs and the preference for simpler models in machine learning, we conclude that using only lagged 1 predictors is sufficient. The models suggested using 9 nearest neighbors as the k hyperparameter.

Table 7: KNN Regression Model Performance Using RMSE

Model	RMSE	K
$\log_{10}(C_t) \sim \log_{10}(N_{t-1} \cdot Flow_{t-1})$	0.0006698	9
$\log_{10}(C_t) \sim \log_{10}(N_{t-1} \cdot Flow_{t-1}) + \log_{10}(N_{t-2} \cdot Flow_{t-2})$	0.0006592	9
$\log_{10}(C_t) \sim \log_{10}(N_{t-1}) * \log_{10}(Flow_{t-1})$	0.0006567	9
$\log_{10}(C_t) \sim \log_{10}(N_{t-1}) * \log_{10}(Flow_{t-1}) + \log_{10}(N_{t-2}) * \log_{10}(Flow_{t-2})$	0.0006565	9

Average RMSE for All Models of each Algorithm

Table 8 shows the average RMSE of all models for each algorithm. We also have evidence that K-Nearest Neighbors Regression has the lowest average error regardless of the models being fit. Hence, we can conclude that KNN is more stable compared to the other models.

Table 8: Average RMSE for Each Algorithm

Algorithm	Average
Linear Regression	0.0007530
Linear Regression with Interaction	0.0007286
Ridge Regression	0.0007327
Generalized Additive Model	0.0007271
Decision Tree Regression	0.0007327
Gradient Boosting Regression	0.0006923
KNN Regression	0.0006606

Conclusion

Our analysis revealed several key findings:

- Best Models: K-Nearest Neighbors Regression (KNN) with 9 nearest neighbors consistently achieved the lowest prediction error and demonstrated robustness, especially when using interaction terms.
- Exclusion of PMMoV: Excluding PMMoV from the models improved performance.
- Normalization vs. Interaction for Flow: Models normalized by flow performed better than unnormalized models. Interaction terms were found to be more effective than normalization.

- Lagged Data: Lagged 1 predictors were sufficient than using both lagged 1 and lagged 2 predictors.
- Linearity: Linear model assumption met

Based on these findings, we selected KNN with k=9 as the ideal model for predicting COVID-19 case counts.

The hyperparameters balance complexity and generalization, ensuring effective learning without overfitting. This study highlights the value of advanced regression techniques and machine learning models, particularly KNN, for wastewater surveillance data. It should be noted that one of the model assumptions of KNN is the even scale of variables, which is met in our study since we only used a single variable. Proper data processing should be applied if other variables are study along with RNA concentration.

Future research should focus on integrating additional data sources and refining models to enhance predictive accuracy and robustness, supporting wastewater-based epidemiology as an early warning system for infectious disease outbreaks.

References

Ahmed, Warish, Nicola Angel, Janette Edson, Kyle Bibby, Aaron Bivins, Jake W O'Brien, Phil M Choi, 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, Sudipti, Aditi Nag, Jasmine Sethi, Jayana Rajvanshi, Sonika Saxena, Sandeep K Shrivastava, and Akhilendra Bhushan Gupta. 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): 2823–36.

Barasa, Edwine W, Paul O Ouma, and Emelda A Okiro. 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, Vinson Wai-Shun, Peter Ka-Fung Chiu, Chi-Hang Yee, Yuhong Yuan, Chi-Fai Ng, and Jeremy Yuen-Chun Teoh. 2021. "A Systematic Review on COVID-19: Urological Manifestations, Viral RNA Detection and Special Considerations in Urological Conditions." *World Journal of Urology* 39 (9): 3127–38.

Chen, Yifei, Liangjun Chen, Qiaoling Deng, Guqin Zhang, Kaisong Wu, Lan Ni, Yibin Yang, et al. 2020. "The Presence of SARS-CoV-2 RNA in the Feces of COVID-19 Patients."

Journal of Medical Virology 92 (7): 833–40.

Cheung, Ka Shing, Ivan FN Hung, Pierre PY Chan, KC Lung, Eugene Tso, Raymond Liu, YY Ng, 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.

Hamza, Hazem, Neveen Magdy Rizk, Mahmoud Afw Gad, and Ibrahim Ahmed Hamza. 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): 2707–13.

Hasan, Shadi W, Yazan Ibrahim, Marianne Daou, Hussein Kannout, Nila Jan, Alvaro Lopes, Habiba Alsafar, and Ahmed F Yousef. 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, Maria, Nicklas Paxéus, Lars Magnius, Lucica Enache, Birgitta Arnholm, Annette Johansson, Tomas Bergström, and Heléne Norder. 2014. "Detection of Pathogenic Viruses in Sewage Provided Early Warnings of Hepatitis a Virus and Norovirus Outbreaks." *Applied and Environmental Microbiology* 80 (21): 6771–81.

Kitajima, Masaaki, Brandon C Iker, Ian L Pepper, and Charles P Gerba. 2014. "Relative Abundance and Treatment Reduction of Viruses During Wastewater Treatment Processes—Identification of Potential Viral Indicators." *Science of the Total Environment* 488: 290–96.

Kitajima, Masaaki, Hannah P Sassi, and Jason R Torrey. 2018. "Pepper Mild Mottle Virus as a Water Quality Indicator." *NPJ Clean Water* 1 (1): 1–9.

Lauer, Stephen A, Kyra H Grantz, Qifang Bi, Forrest K Jones, Qulu Zheng, Hannah R Meredith, Andrew S Azman, Nicholas G Reich, and Justin Lessler. 2020. "The Incubation Period of Coronavirus Disease 2019 (COVID-19) from Publicly Reported Confirmed Cases: Estimation and Application." *Annals of Internal Medicine* 172 (9): 577–82.

Li, Ruiyun, Sen Pei, Bin Chen, Yimeng Song, Tao Zhang, Wan Yang, and Jeffrey Shaman. 2020. "Substantial Undocumented Infection Facilitates the Rapid Dissemination of Novel Coronavirus (SARS-CoV-2)." *Science* 368 (6490): 489–93.

Maal-Bared, Rasha, Yuanyuan Qiu, Qiaozhi Li, Tiejun Gao, Steve E Hrudey, Sudha Bhavanam, Norma J Ruecker, Erik Ellehoj, Bonita E Lee, and Xiaoli Pang. 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, R Handsher, T Halmut, M Neuman, A Bobrov, H Rudich, A Vonsover, L Shulman, O Kew, and E Mendelson. 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): 1670–75.

Mark, Shannon, Mason Carolyn, Stacey Laura, Sara Stephanie, Stephanie Zachary, Tomothy Daniel, and Charles. 2024. Multidisciplinary Digital Publishing Institute. "Wastewater Surveillance of SARS-CoV-2 in Minnesota." *Water* 16 (2024. Multidisciplinary Digital Publishing Institute): 541.

Nishiura, Hiroshi, Tetsuro Kobayashi, Takeshi Miyama, Ayako Suzuki, Sung-mok Jung, Katsuma Hayashi, Ryo Kinoshita, et al. 2020. "Estimation of the Asymptomatic Ratio of Novel Coronavirus Infections (COVID-19)." *International Journal of Infectious Diseases* 94: 154–55.

Oran, Daniel P, and Eric J Topol. 2020. "Prevalence of Asymptomatic SARS-CoV-2 Infection: A Narrative Review." *Annals of Internal Medicine* 173 (5): 362–67.

Parasa, Sravanthi, Madhav Desai, Viveksandeep Thoguluva Chandrasekar, Harsh K Patel, Kevin F Kennedy, Thomas Roesch, Marco Spadaccini, 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): e2011335–35.

Peccia, Jordan, Alessandro Zulli, Doug E Brackney, Nathan D Grubaugh, Edward H Kaplan, Arnau Casanovas-Massana, Albert I Ko, et al. 2020. "Measurement of SARS-CoV-2 RNA in Wastewater Tracks Community Infection Dynamics." *Nature Biotechnology* 38 (10): 1164–67.

Post, Lori Ann, Tariq Ziad Issa, Michael J Boctor, Charles B Moss, Robert L Murphy, Michael G Ison, Chad J Achenbach, 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.

Randazzo, Walter, Pilar Truchado, Enric Cuevas-Ferrando, Pedro Simón, Ana Allende, and Gloria Sánchez. 2020. "SARS-CoV-2 RNA in Wastewater Anticipated COVID-19 Occurrence in a Low Prevalence Area." *Water Research* 181: 115942.

Rosario, Karyna, Erin M Symonds, Christopher Sinigalliano, Jill Stewart, and Mya Breitbart. 2009. "Pepper Mild Mottle Virus as an Indicator of Fecal Pollution." *Applied and Environmental Microbiology* 75 (22): 7261–67.

Weidhaas, Jennifer, Zachary T Aanderud, D Keith Roper, James VanDerslice, Erica Brown Gaddis, Jeff Ostermiller, Ken Hoffman, 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, Martin CS, Junjie Huang, Christopher Lai, Rita Ng, Francis KL Chan, and Paul KS Chan. 2020. "Detection of SARS-CoV-2 RNA in Fecal Specimens of Patients with Confirmed COVID-19: A Meta-Analysis." *Journal of Infection* 81 (2): e31–38.

Ye, Yinyin, Robert M Ellenberg, Katherine E Graham, and Krista R Wigginton. 2016. "Survivability, Partitioning, and Recovery of Enveloped Viruses in Untreated Municipal Wastewater." *Environmental Science & Technology* 50 (10): 5077–85.

Zhan, Qingyu, Kristina M Babler, Mark E Sharkey, Ayaaz Amirali, Cynthia C Beaver, Melinda M Boone, Samuel Comerford, 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, Tao, Mya Breitbart, Wah Heng Lee, Jin-Quan Run, Chia Lin Wei, Shirlena Wee Ling Soh, Martin L Hibberd, Edison T Liu, Forest Rohwer, and Yijun Ruan. 2006. "RNA Viral Community in Human Feces: Prevalence of Plant Pathogenic Viruses." *PLoS Biology* 4 (1): e3.