

# Optimization of correction factors for WBS in SARS-COV-2 uncertainty factors

Wei Huang, Junchao Zhang, Dong Pan\*, Yisheng Huang

Guilin University of Technology AT Nanning, Guangxi, China

\* Corresponding Author Email: 41725065@qq.com

**Abstract.** Effluent-based epidemiological methods based on effluent analysis are a new method to study drug consumption. In this study, CF was reestimated and improved by establishing a reasonable model to improve the accuracy of the WBE method. Before analyzing the obtained data, we need to eliminate errors and their blank data. Then, after downloading the exported data, knowing that there were 1099 sewers in the data downloaded on CDC, the data table was filtered, the blank data was excluded, and the data from April 24, 2022 to May 8, 2022 was selected as the percentage change in SARS-CoV-2RNA levels at 15-day intervals; Publication of SARS-CoV-2 concentration data sheets in wastewater in nuclear-weapon States. The data on CoV-2 concentration in wastewater were screened and eliminated, deleted from the blank data of SARS-CoV-2 virus concentration, and classified according to the measured serial number of sewers, and the SARS-CoV-2 virus concentration data of a total of 451 sewers were summarized. Then select based on the heatmap of the number of new cases in each state in the last 30 days, and based on the heatmap of the number of new cases in each state in the last 30 days, select the abbreviated state name of CA, CT, ID, IL, IL, KS, KY, MA, MI, MO, NE, NV, New York, Ohio, Oregon, TX, UT, VA, WA, WV, WI. Analyze the data. Finally, based on K-means clustering, linear regression and bootstrap algorithms, combined with relevant parameters, normal distribution and bias distribution, a new CF is re-established to improve the accuracy of WBE.

**Keywords:** *wastewater analysis, Epidemiological methods for investigating drug consumption, K-means clustering, bootstrapping.*

## 1. Introduction

It has been two years since the outbreak of novel coronavirus pneumonia in 2019. The novel coronavirus has undergone several mutations and there are now 11 mutated strains, including Alpha, Beta, Delta, and Omicron. Among them, the mutant strain Omicron has been prevalent in many countries and regions in the world. Compared with the previously prevalent mutant strains, Omicron has a large number of key mutations and is more transmissible and insidious, and has immune escape, making it more likely to be distributed at multiple points or have concentrated outbreaks. As of 2021-12-08, Omicron mutants have been reported in 57 countries and regions worldwide, and as of 2022-01-20, Omicron has spread to 14 provinces in China. Although it is possible to rapidly detect and effectively isolate infected individuals through regional full-scale nucleic acid testing, large-scale nucleic acid testing is costly. However, large-scale nucleic acid testing is costly and has a large economic and livelihood impact, so its use is limited. To address this issue, in August 2020, the World Health Organization (WHO) released a scientific brief entitled "Status of environmental surveillance for SARS-CoV-2 virus". This paper summarizes the background, methods, and uses of neo coronavirus surveillance through municipal wastewater, and provides a guide for the global fight against the epidemic. Wastewater monitoring is based on the principles of wastewater- Based Epidemiology (WBE), which analyzes the concentration of contaminants or biomarkers in municipal wastewater treatment plant influent and infer the status of the substance in the wastewater catchment by combining information on human metabolic mechanisms, influent flow and the number of people served. This approach may play an important role in future large-scale prevention and control of new crown outbreaks.

One of the main challenges of the wastewater epidemiology (WBE) approach for new crown outbreak estimation is the uncertainty in the values of a number of variables that arise from sample

collection, substance stability, population size estimation, human metabolic ratio of substances, concentration determination of substances, and influent flow to wastewater plants. The correction factor (CF) is an important parameter for translating the mass of biomarkers measured in the effluent into the final consumption of the population, and the CF is considered to be a source of uncertainty that has a significant impact on the accuracy of the final consumption estimates. However, most of the CF estimates are derived from some old pharmacokinetic studies, so this study proposes to re-estimate and improve the CF by building a reasonable model to improve the accuracy of the WBE method.

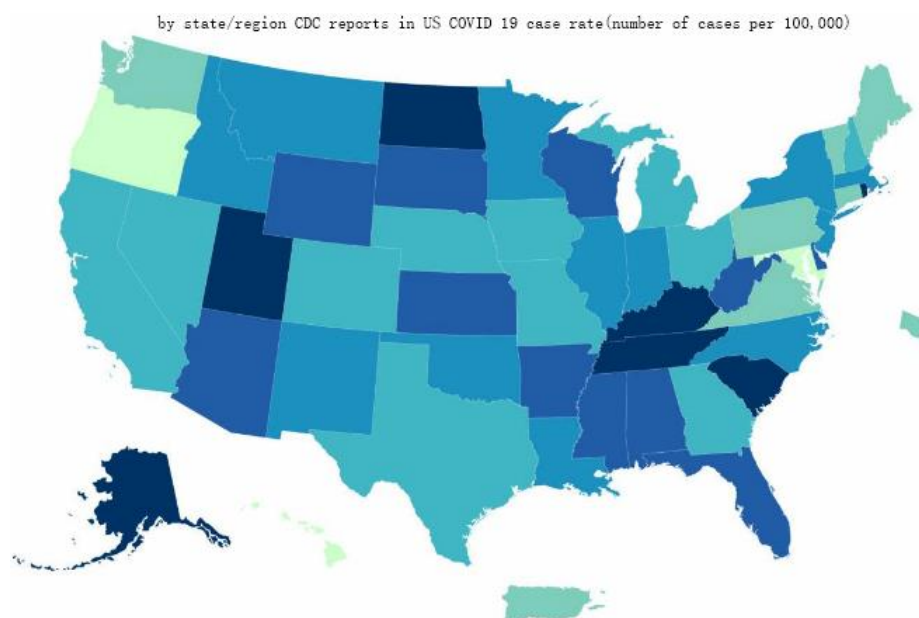
## 2. Model building and solving

### 2.1. Uncertainty: Modeling of sample collection point analysis

People infected with SARS-COV-2 can excrete viral RNA (genetic material from the virus) in their feces, and this RNA can be detected in community wastewater. Wastewater, also known as sewage, includes water from households or buildings (e.g., toilets, showers, and sinks) that may contain human fecal waste, as well as water from non-household sources (e.g., rainwater and industrial uses).

Since we only use the first 15 days of the sewer collection point, we only need to use the data downloaded from the CDC to analyze one dimension of the sewer point, location.

Before analyzing the resulting data, we need to eliminate errors and their gaps. Then, knowing from the downloaded and exported data that the distribution of wastewater monitoring sampling points in the United States is reasonable in order to build a linear regression model to obtain a new correction factor CF1 to optimize the effect of uncertainty in sample collection, we must know the distribution of wastewater monitoring sampling points established in each state. Based on the large amount of data, the K-means algorithm we use is the algorithm proposed by J.B. MacQueen in 1967 [1], which classifies the downloaded 200,000 data into 23 clusters (states), and selects the states with more new cases through the heat map (Figure 1), simplifying the data without affecting the accuracy.

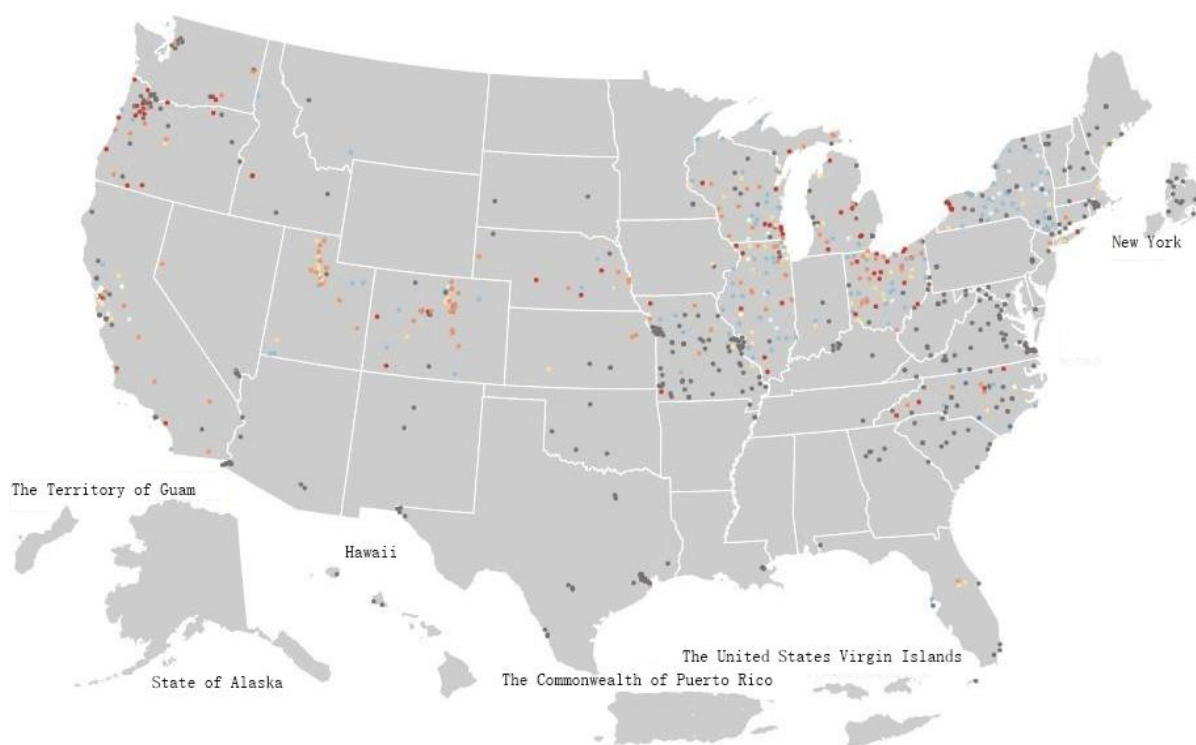


**Figure 1.** Heatmap of new cases by state in the United States

Second, due to the presence of viral particles or genetic material ribonucleic acid (RNA) in the feces/urine of infected individuals, and increasing evidence that intestinal system tissue cells may also be vulnerable to SARS-COV-2 invasion [2-3]; However, intestinal discomfort such as diarrhoea has also been identified as one of the typical symptoms of COVID-19 [4]. This is why the presence of SARS-COV-2 genetic material (RNA) has been repeatedly detected in the feces/urine of infected individuals [5]., the virus has the potential to enter the water environment (sewage) through the drainage pipe system [6-7], so Dutch scholars have proposed the idea of monitoring the emergence

and spread of SARS-COV-2 (COVID-19-induced virus) based on wastewater epidemiology (WBE). At present, the detection of viruses or their genetic material RNA from sewage also provides a new idea for virus transmission warning or outbreak judgment: SARS-COV-2 genetic material RNA can be detected in sewage to determine the presence and number of people infected with COVID-19, feasible viral early warning mechanisms can be established based on sewage epidemiology (WBE) or environmental testing (ES) diseases [8-9]. The feasibility of this technology as an early warning of outbreaks has been demonstrated and can characterize the dynamics and trends of virus transmission, estimate the proportion of infected groups, and even track infected groups within a certain range based on WBE results. Therefore, we only need to cluster aggregates that have significantly changed their SARS-COV-2RNA levels at sewage collection sites in the last 23 days within the 15-day intervals defined by "date\_start" and "date\_end", as well as the location of new coronavirus cases in the United States in recent days, close to sewage collection sites, to understand whether the distribution of sampling sites is reasonable.

During the same period in the previous 15 days, we selected different counties to analyze the distribution of sewage sampling points. First, we learned from the CDC Coronavirus Data Tracker (CDC.gov) that the distribution of sewage collection points across the United States is shown in the graph. Arabic numeral.



**Figure 2.** Map of U.S. National Wastewater Collection Points from April 16, 2022 to May 8, 2022

From the distribution chart, we can see that the distribution of all collection points is scattered, some are closely clustered, some are scattered clusters in some states, and some gray points are blank, so we have to do data preprocessing, use EXCEL filtering tools to remove gray point-blank data, and then we manually select each of the nine states for sewage collection point regression analysis. using EXCEL filtering tool to eliminate the gray points blank data, and then we manually select each of the nine states sewage collection points for regression analysis. The filtering tool of EXCEL was used to eliminate the gray point-blank data and select the locations of counties with denser ( $>10,000$  people) population data from April 24, 2022 to May 8, 2022 to filter nearly 379 data from nearly 200,000 See supporting materials. The data were then analyzed using the K-means algorithm and the linear regression algorithm.

## 2.2. Model solving

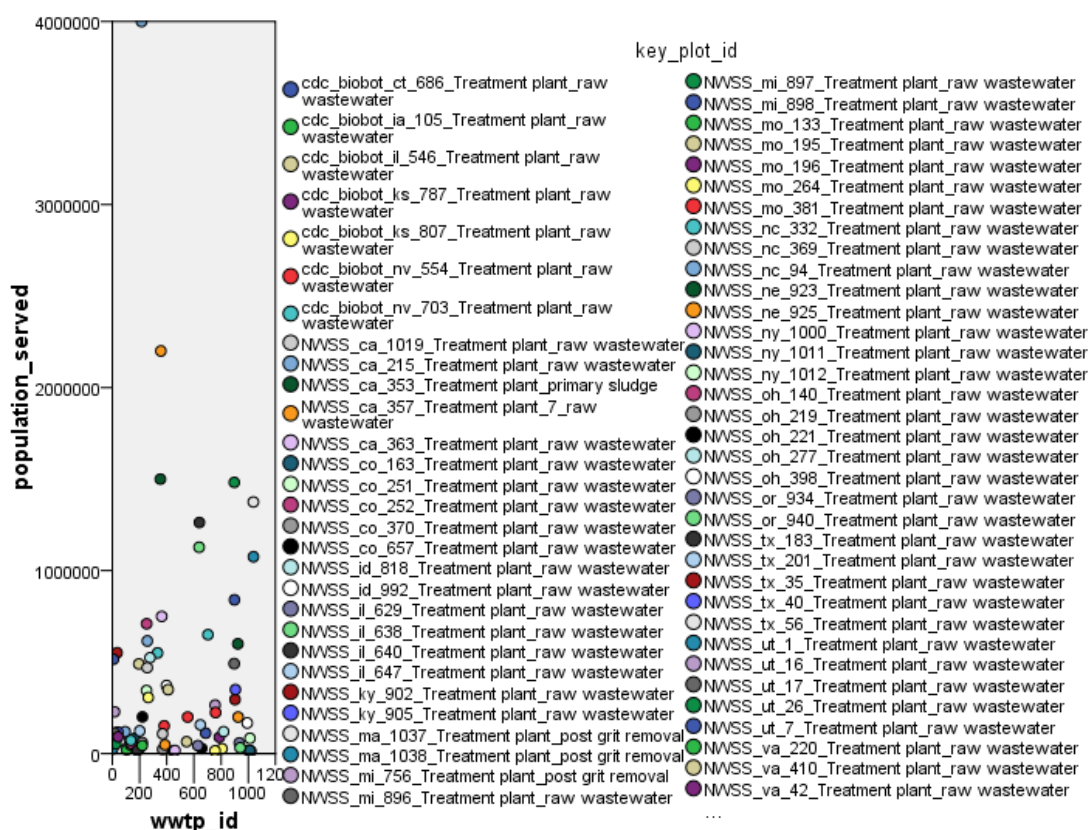
To ensure a more accurate understanding of the distribution of sewage collection points in the United States, we addressed different spatial issues at selected times. We chose to analyze daily analysis of the distribution of sewage collection sites and the percentage change in SARS-COV-2 over the previous 15 days in 23 states.

### 2.2.1. Cluster analysis

Due to the large amount of data, we used "sewage collection points", "percentage change" and "county population" as clustering criteria to divide the data into 23 categories, the clustering process is shown in the table, and the other 22 clusters are clustered in the table shown in the supporting materials (Table 1). The clustering results are shown in Figure 3. As shown, we divided the 190,000 data points into 23 zones, each with 2, 3, and 5 sewer points, for a total of 72 sewer points.

**Table 1.** Clustering results

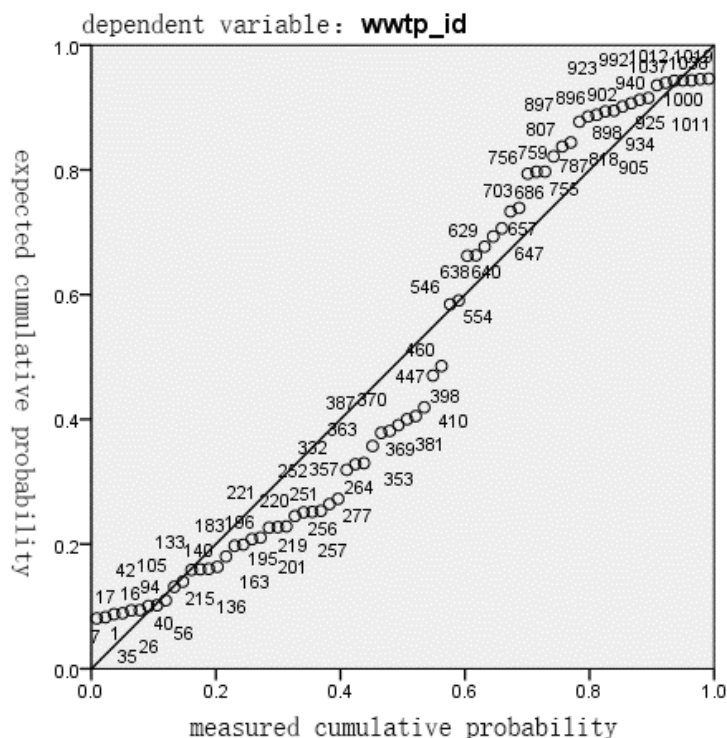
|                   | 1       | 2      | 3     | 4      | 5      |
|-------------------|---------|--------|-------|--------|--------|
| population_served | 1482000 | 840600 | 49640 | 492000 | 263057 |
| ptc_15d           | -100    | -100   | 556   | -94    | 629    |



**Figure 3.** Clustering charts

### 2.2.2. Linear regression model solving

The new coronavirus data tracker on the CDC website was downloaded and analyzed the rate of change in SARS-COV-2RNA in state sewers based on SARS-COV-2RNA wastewater indicator data from the last 15 days. The regression coefficient can obtain the constant term 468.185 and the regression coefficient  $1.937E-7$ , respectively, and the linear regression equation is  $y=468.185+1.937E-7x$ ; Since the significance  $P$  (sig)  $0.086 > 0.01$  so the linear regression coefficient is more significant, and the regression model is obtained as:  $y = 468.185 + 1.937E-7x$  gives the new CF1 correction factor of  $1.937E-7$ . A normal P-P plot of the regression standardization was made using SPSS software (Fig. 4) to determine a better fit for the new correction factor CF1.



**Figure 4.** Normalized residual normal plot

### 2.2.3. Model testing

The significance test of linear regression equations is usually performed by the analysis of variance (ANOVA) method F test, which uses the principle that the sum of squares can be decomposed into the sum of squares and the sum of squares of regression. The significance of the linear regression equation  $y=468.185+1.937E-7x$  is tested to obtain Table 2.

**Table 2.** Analysis of variance Table 1

|            | Sum of squares | interrupter | Mean square | F     | p             |
|------------|----------------|-------------|-------------|-------|---------------|
| regression | 336307.719     | 1           | 336307.719  | 3.035 | 008.6 million |
| remnants   | 7757406.059    | 70          | 110820.087  |       |               |
| total      | 8093713.778    | 71          |             |       |               |

Note: a. Dependent variable: wwtp\_id; b. Predictors: (constant), ptc\_15d

We know  $SSR=336307.719$ ,  $SSE=7757406.059$ ,  $F=3.035$ ,  $P=0.086>0.05$ . The significance of the S o regression equation is the mean. In fact, the significance of the regression equation is equal to the test result of the standardized regression coefficient for a univariate linear regression problem.

## 2.3. Uncertainty: An analytical model used to estimate population size

### 2.3.1. Linear regression model

It is required to make the coverage as high as possible while considering factors such as cost according to the characteristics of base station coverage. Firstly, the base station coverage is determined according to the characteristics of BTS coverage, the points not covered by the existing base station are selected, and the coordinates of the base station establishment and BTS establishment types selected in the first question are used to simulate different main direction angles with the maximum coverage as the goal, and the corresponding base station coordinates and base station main direction angles are derived by solving.

The CDC-derived data for the last 7 days of new coronavirus cases show that the number of confirmed new coronavirus cases on the 7th day of the year in states with a high increase in new coronavirus in the United States may be correlated with the percentage of SARS-COV-2 collected at sewage collection points on day 15 of this year and several recent independent variables in the state.



Sewage RNA testing in the Netherlands showed a positive sewage RNA test result in the A mersfoort area on March 5, 2020, but there were no clinically confirmed cases of COVID-19 in the region at that time, which proved the feasibility of W BE as an early warning sign of the virus [10]. Therefore, we used linear regression to analyze the relationship between them and combine the principles of sewage epidemiology to predict the number of new outbreaks that will continue to occur in these states in the future. This sewer epidemiology has the advantage of being widespread and wide, allowing for an assessment of overall infection in the wider community. Municipal sewage treatment plants receive sewage from thousands of urban households through a network of sewers, and the population of sewage collection areas can range from thousands to millions of people, covering a wide range of people, including those who are not infected in time or do not find symptoms without nucleic acid testing, thereby improving the accuracy of the overall size of the uncertainty factor (number of people served) of the WBS through regression The regression coefficient of the analytical model is the latest correction factor.

Principle: Linear regression is used to characterize the degree to which changes in single or multiple values affect a numeric variable.

### 2.3.2. Solve the model

The regression coefficient can obtain the constant term 472.107 and the regression coefficient 2.039E-5, respectively, and the linear regression equation is  $y=472.107+2.039E-5x$ ; due to the significance  $P(\text{sig})0.086>0.01$  so the linear regression coefficient is more significant, and the regression model is:  $y = 472.107 + 2.039E-5x$  gives the new CF<sub>2</sub> correction factor of 2.039E-5.

A normal P-P plot of the regression standardization was made using SPSS software to determine a better fit for the new correction factor CF<sub>2</sub> (Fig. 5).

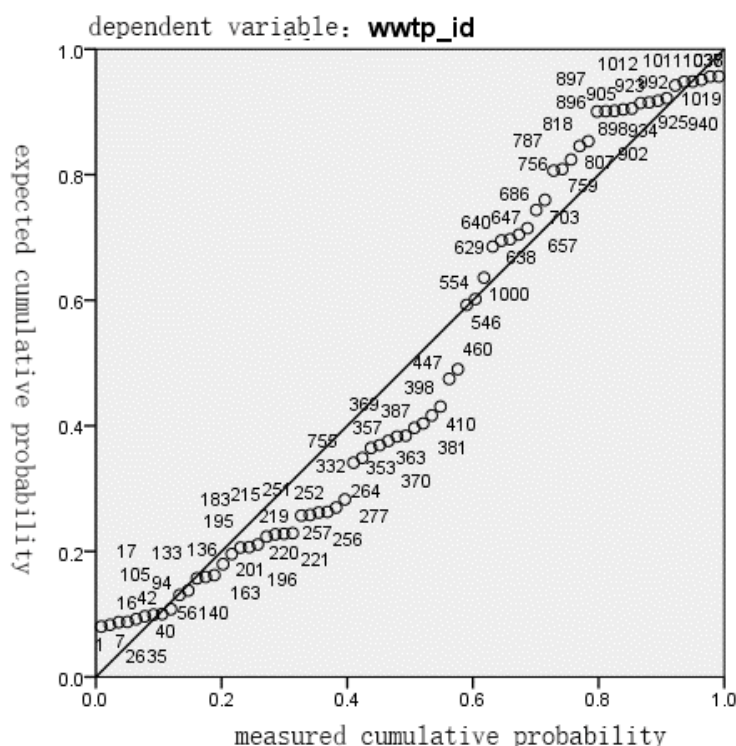


Figure 5. Normal P-P plot of regression standardized residuals

### 2.3.3. Testing of the model

The linear regression equation  $y=472.107+2.039E-5x$  was tested for significance and Table 3 was obtained.

**Table 3. ANOVA Table 2**

|            | Sum of squares | interrupter | Mean square | F     | p           |
|------------|----------------|-------------|-------------|-------|-------------|
| Regression | 11042.901      | 1           | 11042.901   | 0.096 | 0.758 bytes |
| Remnants   | 8082670.877    | 70          | 115466.727  |       |             |
| Total      | 8093713.778    | 71          |             |       |             |

The regression equation is important because SSR=11042.901, SSE=8082670.877, F=0.096, P=0.758>0.05, so the regression equation is significant. In fact, the significance of the regression equation is equal to the test result of the standardized regression coefficient for a univariate linear regression problem.

## 2.4. Uncertainty: An analytical model used to determine the ratio of human metabolism to the concentration of a substance

### 2.4.1. Data preprocessing

People infected with SARS-COV-2 can excrete viral RNA (genetic material from the virus) in their feces, and this RNA can be detected in community wastewater. Wastewater, also known as sewage, includes water from homes or buildings (e.g., toilets, showers, and sinks) that may contain human fecal waste, as well as water from non-household sources (e.g., rainwater and industrial uses).

- Wastewater from the sewer is collected as it flows into the treatment plant.
- These samples were sent to an environmental or public health laboratory for SARS-COV-2 testing.
- The Department of Health submits test data to CDC through the online NWSS Public Health Incident Response Data Collation and Integration (DCIPHER) portal.
- The NWSSDCIPHER system analyzes the data and reports the results to the health department for use in its COVID-19 response.

Since we only used the sewer collection points for the 15 days before May 3, 2022, we only used the downloaded data to analyze the sewer points in one dimension, the sewer outlet number.

Based on the NWSSPublicSARS-CoV-2 concentration data in wastewater and SARS-CoV-2\_wastewater indicator data from the supporting materials, ptc\_15d was selected as the SARS-CoV-2 concentration and pcr\_conc\_smoothed as the smoothed sample bar normalized to the mobile population or human feces to fit the SARS-CoV-2 virus. concentrations. The abbreviations of states ND, UT, KY, TN, SC, KS, SD, AZ, and NY with a high number of new cases in the heat map were filtered by EXCEL data to exclude the blank data of ptc\_15d and pcr\_conc\_smoothed, and recreated into concentration table for analysis.

### 2.4.2. Model building

Meaning of normalized indicators according to the derived data.

Normalization method for the concentration of SARS-CoV-2 virus in wastewater. There are two normalizations: mobile population or human feces. The flow-population normalization method is to multiply the SARS-CoV-2 concentration and the flow rate and then divide by the service population. The flow-population normalization method indicates whether the total number of individuals being shed at the site has changed. Human fecal normalization is the SARS-CoV-2 virus concentration divided by the human fecal marker concentration. The human fecal normalization method indicates whether the proportion of individuals shedding at the site has changed.

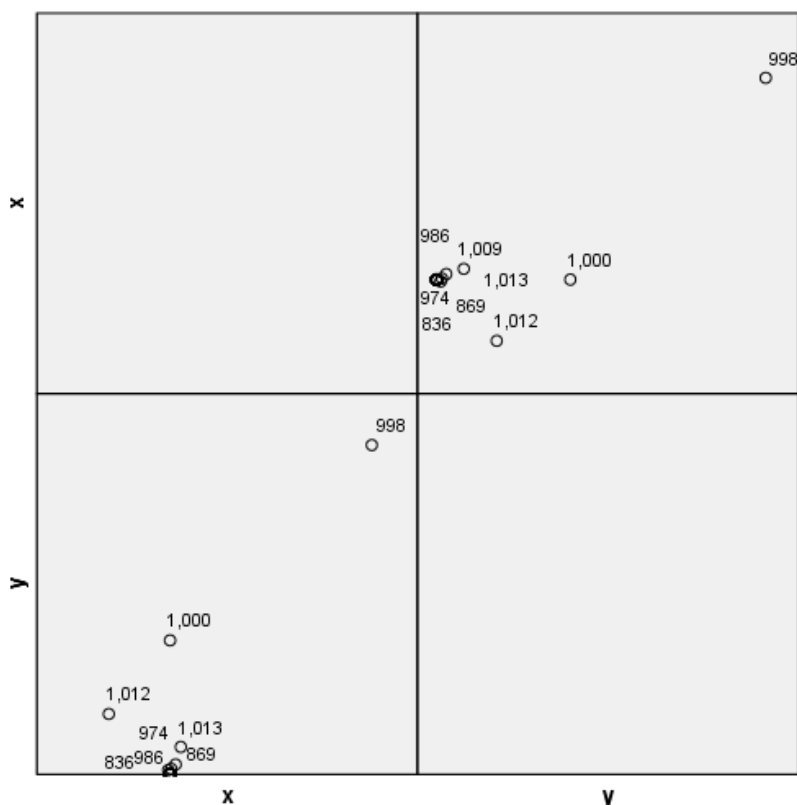
The human content of investigated SARS-COV-2RNA is generally calculated by multiplying the measured SARS-COV-2RNA concentration by the final flux and the corresponding correction factor (CF3), and then dividing by the population size to obtain the continental per capita content.

In summary, a linear regression model is constructed.  
population served

$$* m1 = CT * F * CF3 \quad (1)$$

### 2.4.3. Solving of the model

The linear regression model is solved by the same model solution as above, and the regression coefficients can be obtained as the constant term 53407792760000 and the regression coefficient 57.860 ( $CF_3$ ) respectively, thus the linear regression equation is  $y=53407792760000+57.860x$ ; due to the significance  $P(\text{sig})0.08>0.01$  so the linear regression coefficient is more significant, the regression model is derived as:  $y=53407792760000+57.860x$ . The correlation matrix is shown in Fig. 6.



**Figure 6.** Correlation matrix

### 2.4.4. Testing of the model

Significance test of linear regression equation  $y=53407792760000+57.860x$  was performed to obtain Table 4.

**Table 4.** ANOVA Table 3

|            | Sum of squares | interrupter | Mean square | F       | p                  |
|------------|----------------|-------------|-------------|---------|--------------------|
| regression | 4.5326E30      | 1           | 4.5326E30   | 102.340 | 0.000 <sup>b</sup> |
| remnants   | 2.0816E30      | 47          | 4.429E28    |         |                    |
| total      | 6.6143E30      | 48          |             |         |                    |

We know:  $SSR=4.5326E30$ ,  $SSE=2.0816E30$ ,  $F=102.340$ ,  $P=0<0.05$ , so the regression equation is significant, in fact the significance of the regression equation is equal to the test result of the standardized regression coefficient for the univariate linear regression problem.

## 3. Model evaluation

### 3.1. Advantages of the model

The model clusters more than 200,000 data points in the raw data, simplifying the data while ensuring the reliability of the solution

This model analyzes the distribution of sewer inlet collection points in a given temporal and spatial dimension, giving us more accurate results.



The model's more satisfactory solution is better able to determine whether the distribution of sewer collection points in the United States is reasonable and reluctantly produces new and more reasonable correction factors.

The model simplifies the problem.

### **3.1.1. Disadvantages of the model**

Because there is little data taken from individual county indicators, the model cannot analyze and determine correction factors more accurately

There are still uncertainties in the model.

The model has some large biases.

## **4. Suggestion**

### **4.1. Early warning recommendations for companies, treatment plants and researchers**

Traditional "point-to-point" and "interventional" nucleic acid and antibody detection screening methods are not only laborious and laborious, but also cause errors or omissions due to the different stages of individual infection, making it difficult to screen, isolate and control outbreaks in advance. This makes screening, isolating and controlling outbreaks difficult. On March 24, 2020, the RNA of SARS-CoV-2 was detected in sewage for the first time in the Netherlands, and a growing body of research has shown that the virus can enter the intestinal system, and the application of WBE provides timely information for outbreak warning. The application of WBE provides timely information for the early warning of emerging outbreaks, and the analysis of the genetic material content of SARS-CoV-2 in sewage can analyze the number of clinically confirmed cases in an area, providing strong evidence of virus transmission trends. However, the WBE's response to transmission dynamics is hampered because people with COVID-19 infection can still test positive in their stool even after recovery. From another perspective, it makes more sense to construct a plan that reflects the downward trend of the virus than to detect the upward trend of the virus, which will provide more accurate information for the prevention and control of the epidemic. In order to strengthen epidemic prevention and control, more attention should be paid to the sensitivity of low-concentration RNA detection techniques, quantitative data of RNA in sewage, and the accuracy of results using composite samples with varying flow rates [11]. Support and encourage research teams to cooperate with local sewage treatment plants, companies, etc. in WBE projects to predict epidemic prevention and control.

### **4.2. Government recommendations for epidemic prevention and control**

First, the social governance of migrants is crucial. The government should attach great importance to the risk of epidemic transmission caused by the floating population, and do a good job in the management of the floating population and the early warning of the epidemic. Most parts of the United States have zero new confirmed cases for several days, and epidemic prevention and control has begun to bear fruit, and the focus of epidemic prevention and control has shifted to preventing imported cases. It is necessary to use big data and other technologies to strengthen the prevention and control of the inflow of overseas personnel and prevent the epidemic from entering from abroad. Second, the government should disclose epidemic information in a timely and accurate manner. The slow reporting, concealment, and omission of epidemic information may delay the best time for epidemic prevention and control, which is not conducive to epidemic prevention and control. Third, we should pay attention to the role of information dissemination efficiency. On the basis of compliance with the law and not believing rumors, the dissemination of information about the epidemic should be increased. The government should also attach importance to the construction of official media platforms, timely disclose authoritative information, and reduce information asymmetry between the government and the public.

## 5. Conclusion

This work was financially supported by the XXX Fund.

## References

- [1] McQueen J. Some classification and analysis methods for multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 1967: 281 - 297.
- [2] LAMERSM M, BERMAN J, VAN DER VAART J, ET AL. SARS-CoV 2 can effectively infect human bowel cancer [J]. Science, 2020, 369 (6499): 50 - 54.)
- [3] XIAO Feng, TANG Mingmingzheng Xiaobin, et al. Evidence of SARS gastrointestinal infection-CoV-2 [J]. Gastroenterology, 2020, 158 (6): 1831 - 1833.)
- [4] Dochertia Harrison E, Green 20 An observational cohort study of 133 UK COVID-19 patients using the ISARIC WHO clinical profiling protocol [J]. British Medical Journal, 2020, 369: m1985.
- [5] Wang De W.HU B.HU C.et Clinical features of 138 patients hospitalized in Wuhan with novel coronavirus pneumonia in 2019 [J]. Journal of the American Medical Association, 2020, 323 (11I.): 1061 - 1069).
- [6] ZHANG Guocheng, HONG YiF CHEN Pengpeng, et al. Gastrointestinal manifestations of SARS-CoV-2 infection and viral load in stool samples in Hong Kong cohort: a systematic review and meta-analysis [J]. Gastroenterology, 2020, 159 (1): 81 - 95).
- [7] Mederma G, Heinen L, Erzinga G, et al. Presence of SARS coronavirus 2 RNA in sewage and its correlation with the prevalence of COVID-19 in the early epidemic in Etheric areas [J]. Environmental Science and Technology Papers, 2020, 7 (7): 511 - 516.
- [8] ZHANG Deyong, Ling HB, HUANG Xia et al. Purulent OL Fangcang Hospital has severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus RNA, potential transmission risk of medical waste and disinfection challenges [J], Total Electron Radiation Science, 2020, 741: 140445.
- [9] Epidemiology based on wastewater: current situation and prospects [J]. Journal of Environmental Science and Health, 2019, 9: 77 - 84.
- [10] Croyce S. Ja, Ahmed Mi, Ahmed Zhu and others. Sensitivity assessment of polio environmental surveillance system [J]. GLPI, 2018.
- [11] Dong Xiang, Quantitative evaluation method of sewage epidemiological uncertainty and variability based on probability and statistics [J]. 2018.