PM566 Final Project

chen wei

Covid-19 new cases can be prediced with the virus concentration in wastewater

abstract

During this worldwide pandemic, most of the policy was based on the assumption that Coronavirus disease can be transmitted by air, the incidence cases can be controlled by strategies to block airborne transmission such as wearing facial masks, keeping social distance, quarantine etc. While prevention is necessary, the virus concentration in air is not consistent. Based on the data collected on an official website, there is information that suggest we should examine our assumptions such as Covid19 virus can also survive in water, and monitoring wastewater for the presence of infectious pathogens is a standard practice in public health. During the COVID-19 pandemic, it has been used for detection and quantification of SARS-CoV-2 virus shed into wastewater via feces of infected persons.I hypothised that concentration of the Coronavirus in wastewater can be used as a predictor of incident case numbers of infected people. I found there is some evidence that the concentration of virus in wastewater I support my hypothesis in 19 counties in California. However due to various reason below, further explornary analysis is needed: 1. Most of the waterwaste data is gathered from samples in public places such as universities, restaurants etc. 2. SARS-CoV-2 virus's activity is influenced by waterflow and temperature

Research Question

Can we predict incident case numbers from studies of the virus in the wastewater?

background

In late December 2019, people in Wuhan, China began to get sick with a previously unknown pneumonia, marking the beginning of a new infectious disease, later identified as a new type of coronavirus.Soon this infectious disease became a worldwide pademic. The International Committee on the Taxonomy of Viruses selected the name severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) to represent that the new virus is a sister of the original SARS virus. The disease the virus causes was named coronavirus disease 2019 (COVID-19) by the World Health Organization (WHO). The situation continues to change rapidly. multiple policy such as 6 feet social distancing, facial masks, and quaratine are all based on the SARS-CoV-2 can be transmitted by airbone. but there are few research to see that if SARS-CoV-2 can be transmitted by water. During the COVID-19 pandemic, it has been used for the detection and quantification of SARS-CoV-2 virus shed into wastewater via feces of infected persons. Wastewater surveillance tracks ""pooled samples"" that reflect the overall disease activity for a community serviced by the wastewater treatment plant (an area known as a ""sewershed""), rather than tracking samples from individual people. Collecting and analyzing wastewater samples for the overall amount of SARS-CoV-2 viral particles present can help inform public health about the level of viral transmission within a community. Data from wastewater testing are not intended to replace existing COVID-19 surveillance systems, but are meant to complement them. While wastewater surveillance cannot determine the exact number of infected persons in the area being monitored, it can provide the overall trend of virus concentration within that community.

Method

Wastewater collection

Wastewater from communities is collected by wastewater systems and transported to wastewater treatment plants. Participating utilities collect samples of untreated wastewater or primary sludge. These samples are sent to environmental laboratories for SARS-CoV-2 testing. The testing data, along with the associated utility metadata, are submitted to participating STLT health departments. Health departments submit these data to CDC through the online NWSS DCIPHER portal. CDC analyzes the data in real time and

reports results to the health department for use in their COVID-19 response. CDC also summarizes the national data on COVID Data Tracker. SARS-CoV-2 RNA is quantified using PCR technology: reverse transcription quantitative PCR (RT-qPCR), reverse transcription digital PCR (RT-dPCR), or reverse transcription droplet digital PCR (RT-ddPCR). Laboratory staff should convert concentration estimates produced by PCR software (in units of copies per reaction or copies per reaction volume) to virus concentrations per volume of unconcentrated wastewater or sludge sample. This conversion accounts for the volume of template used in the PCR (and reverse transcriptase reaction if separate), the concentration factor of nucleic acid extraction, and sample concentration processes. the data was download as .csv from the website https://data.ca.gov/dataset/covid-19-wastewater-surveillance-data-california

Data analysis

To interpret SARS-CoV-2 wastewater measurements, polymerase chain reaction (PCR)-based measurements must be converted to sample concentrations and adjusted for testing and wastewater factors, which may change from sample to sample within a wastewater system and between wastewater systems. PCR measurements must be converted to wastewater concentrations prior to submitting data to NWSS. Viral recovery and fecal normalization will be evaluated by the NWSS analytic engine as described below.

California case

The "Total Tests" and "Positive Tests" columns show totals based on the collection date. There is a lag between when a specimen is collected and when it is reported in this dataset. As a result, the most recent dates on the table will temporarily show NONE in the "Total Tests" and "Positive Tests" columns. This should not be interpreted as no tests being conducted on these dates. Instead, these values will be updated with the number of tests conducted as data is received. When comparing to the state dashboard, the dashboard numbers will correspond to the "Reported" columns in this table. Also note that the dashboard displays data that is one day prior to today's date. https://data.ca.gov/dataset/covid-19-time-series-metrics-by-county-and-state

Software use

the dataset was downloaded as .csv form. Excel was using for slightly adjusted few columns. Most of the data analyzing, figure generating, and data wrangling was completed by R version 4.2.1 (2022-06-23 ucrt).

Statistic analysis

First we download both dataset from the resource website.The data included the incidence casess and cummulative case in 59 county from California started from 2020-2-1, the last updated is 2022-10-11. There are 60085 observation based on date,985 observation in each 59 county in the whole california, which is much bigger than the data download from waterwaste website. The data downloaded from the waterwaste website contains 12773 observation started from 2020-3-19, last updated is 2022-10-18 based on the data include only 21 county. After an basic explornary analysis, 985 observation from case data and 14 observation from waterwaste data was removed.also, the new cases column marked as "NA" is also removed from the case data. after the first step, then lining with the two data set, all case reports prior to start of wastewater reporting was removed.keeping the observations those wastewater and incidence case shared the same county and date at the same time. Finally, there are 15181 observation in the case data was filtered out and applied into statistical analysis.

exploratory data analysis

First, i used the 15181 cases observation and 12759 waterwaste observation. i created two scatterplot seprately based on the case and waterwaste through the date so i can compare them each other.

Showcasing plots

Tab 1

Tab 2

based on the two figure shown above, based on the relatively raw data. we can see that the incidence case and wastewater shown a relatively similar pattern: 1.Los Anglese, Orange County and San Diego share a highest incidence number,which is the most outlines side. Comparing with the wastewater scatterplot, the Los Anglese and Orange County also have the highest virus concentration in wastewater. 2.There are two peak according to the date. First one is around Jan 2021. Los Angeles's wastewater virus concentration and incidence cases are all reached to a peak. Second one is that around Jan 2022, the incidence number in Los angeles reached to another peak, but in the same time, the orange counties virus concnentration in wastewater reached to peak. ## TABLE in a exploranary analysis, we can see a similar pattern between the wastewater virus concnentration and the incidence case number according to there date. But due to the uneven obervation in different county, and uneven interval date, the two figure above is not provide enough conving evidence. So due to the question above, i created 2 summarized table, TABLE1 is based on the average of that waterwaste concentration and incidence number based on counties and date in the same time from 20203-19 to 2022-10-11. TABLE 2 is based on the average of that waterwaste concentration and incidence number based on counties and mark down their latitude and longtitude. ## Showcasing plots {.tabset} ### Tab 1

Tab 2

LINE PLOT

TAB 1

on the first lineplot,even though averaging out the date and counies, there is still two peak in the whole obervation around Jan 2021,Jan 2022 in Los Angeleswhich is similar with the explonary analysis. but when i was trying to find out the association between the virus concentration in wastewater and incidence case by scatterplot, the r-sqaure is weak.

`in the last part, i average out the observation only based on counties and use a leaflet to find out the association`

Scatterplot

TAB 1

from the date, there is a problem that the virus has a periodly activity. it may also explain the peak around the January in 2021 and 2022, so i averaged out each counties. based on the figure above, we can see from the leafplot that the virus and incidence case have a similar distribution,yellow on the Los Angleles part. and also the scatterplot shows a realtively positively distribution. # result for easier comparision and lining up, only shared the same county names and date between 2 data set. After a couple filter and selection, only 15181 observation from case dataset was remained abd 12759 observation from waterwaste was remained. Based on the uneven observation between each counties and date, the mean of new cases(unit:people) and the pcr_target concentration(unit:copied/L) grouped by each date and counties. This would be the 12759 observations basic data which would be analyzed in the county and date dimension . In this data started at 2020-3-19, last updated on 2022-10-18. include 19 big counti around California, we can see that the new cases has a relatively positive trend when the date increased, and the pcr concentration in the waterwaste was also increased. (full data can be seen on appendix.) Then the dataset was calculated grouped by county first. In

Figure 1, the map of California. Figure 1 shown a comparison distribution map of pcr_concentration mean of the virus and the new cases coming out in thie 2 years dataset. In Figure 1, Tthe pcr concentration and new cases are all shows a high number(lighther color) around Los Anglese area. Figure 2 shows a group of lineplot of time series vs pcr_concentration and new cases. We can see that there is all have a peak on July, 2020 in both variable, and a peak on the Janurary 2022. based on this two dimension, there is an similar association between pcr concentration and also the mean.

discussion

reflection

even though there has a relatively obvious shared association between this 2 viriable. But there is still limitations. eventhough the concentration of virus in the wastewater are a realtively strong evidence, but there based on the water survilence resource shown that there is still limitation on detection and quantification for using it as a dataset. It is not possible to reliably and accurately predict the total number of infected individals in a community based on sewage surveillance alone.Wastewater surveillance will not represent homes on septic-based systems.Community-level wastewater surveillance at a wastewater treatment plant will not represent communities or facilities served by decentralized systems, such as prisons, universities, or hospitals that treat their own waste.Low levels of infection in a community may not be captured by sewage surveillance if the quantity of SARS-CoV-2 falls below the limit of detection for lab analysis. except for the data itself, the limit area of dataset also affect the accurately.There is only 19 counties and 2 years of unconesecutive data, which is limited and easily influenced by other factors. # conclusion based on the Figure shown above, there is a similar association between the pcr-concentration from wastewater and also the new daily cases. pcr-concentration from wastewater may can be used as a predictor of covid_19 but still further research is needed.

Dowload place

click downgit

resource

1. https://data.ca.gov/dataset/covid-19-wastewater-surveillance-data-california
2. https://data.ca.gov/dataset/covid-19-time-series-metrics-by-county-and-state
3. https://transition.fcc.gov/oet/info/maps/census/fips/fips.txt
4. https://www.latlong.net/country/united-states-236.html
5. https://www.cdph.ca.gov/Programs/CID/DCDC/Pages/COVID-19/Wastewater-Surveillance.aspx ## {-}