 doi:10.20944/preprints202208.0209.v1

Article

Real-Time and Historical Viral Variant Tracking

James Labadorf [1]

[1] University Hospital System of Cleveland, Population Health 1; james.labadorf@uhhospitals.org

Abstract: Viral variant analysis is a bedrock of the disease surveillance. When combined with temporospatial analysis variant analysis can further the knowledge of disease spread in a study area. This paper suggests a method to perform the analysis in an operational setting which will allow for real-time surveillance of viral variants and allow local public health professionals to rapidly respond to changes in the evolution of the disease. This method includes three main subprocesses: preprocessing, analysis, and rendering. This method can be performed across multiple software platforms. A use case is given in which it was found that this method helped a hospital system understand the spread of SARS-CoV-2 in Northeast, Ohio.

Keywords: cluster analysis; SARS-CoV-2; Variant

1. Introduction

In 2019, the first case of SARS-CoV-2 was identified in Wuhan, China[1]. SARS-CoV-2 quickly spread and entered the United States just one month later in January, 2020[2]. SARS-CoV-2 mutations have been a concern since the beginning of the pandemic with Alpha, Beta, and Gamma variants being identified in the first month of the pandemic. Over the course of the pandemic, the Centers for Disease Control and Prevention would classify a total of eight variants as variants being monitored. When mutations happen, they are troubling because of the new variant will have different characteristics which may include potency, transmission rate, and resistance to medical treatments[3].

While the SARS-CoV-2 raised the public's awareness of viral variants, variant tracking has been a bedrock of virus surveillance [4,5] and vaccine development. In vaccine development, tracking variants is used to identify strains that may be resistant to current vaccines[6].

Another bedrock tool of the infectious disease surveillance is cluster analysis. Cluster analysis involves examining the spatial correlation between different cases. A popular method is hotspot analysis. Hotspot analysis identifies zones where either there is more cases than otherwise expected (called a hotspot) or a zone where the infection rate is lower than expected (called a cold zone).

While it has been recognized that hotspot analysis is useful for viral variant tracking[7–9], there is currently no standard method that is designed specifically for the operational setting like a hospital or public health department. As opposed to the research environment, operational settings require reviewing the data at a high frequency in order to intervene, usually at the cost of statistical significance. Having a real-time variant analysis report will help public health professionals in deploying resources, structuring interventions, and understanding the progression of the disease in their study area.

In this paper, a method is detailed which can be deployed inside an operational setting for real-time variant tracking and analysis. Additionally, a narrative of how this method was deployed inside University Hospital System of Cleveland (University Hospitals) is provided.

2. Method

The task of creating cluster analyses has three subprocesses: preprocessing, analysis, and rendering. Each of these subprocesses has a unique and necessary objective. The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 August 2022 doi:10.20944/preprints202208.0209.v1

aim of the preprocessing subprocess is to extract, transform, and load the data into the analysis software. Once complete the analysis subprocess performs any required analysis and returns a data file which then is displayed by the rendering subprocess.

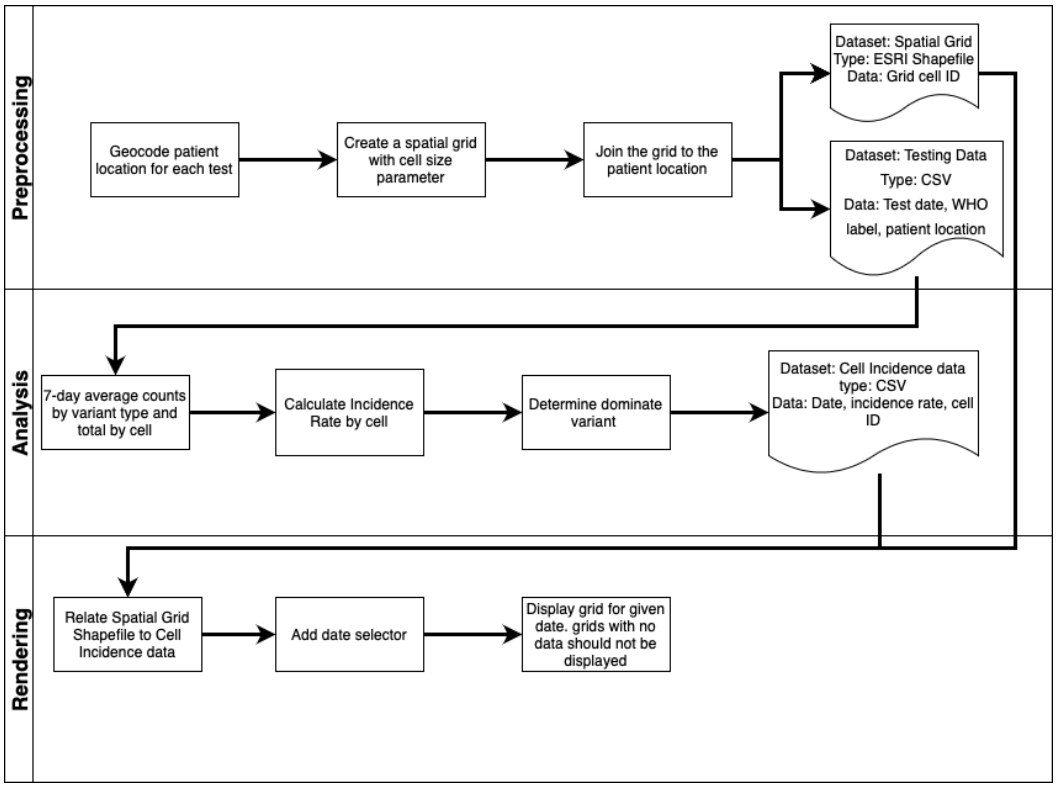

Figure 1. a diagram of the method workflow

2.1. Preprocessing

the output of this subprocess is two files, one with a spatial grid of the extent of the study area; the second is a count of the number of each variant by grid cell ID and date. The exact nature of this subprocess will vary based on the use case and organizational procedure. However, most locations will be able to provide a tabular file with a patient identifier, data of test, and variant label. With the patient identifier, the patient residential address can be extracted from electronic records, geocoded, and then spatially joined to the spatial grid (which will be discussed later in this section).

A spatial grid needs to be generated to cover the study area. Before generating the grid, the spatial extent needs to be set either by crafting a bounding box or using a spatial file with the extent already defined, if the software application in use allows setting the extent. Additionally, a grid size parameter should be set. In setting this parameter it is important to consider both the density of the cases and the underlying geography. A few sample analyses may be necessary to tune this parameter. Once the extent and cell size is defined the grid can be generated. This grid should consist of polygons and not be in raster form. It should have an identifier for each cell.

The next step in the preprocessing subprocess is to join the spatial grid to the case file. The goal here is to insert the cell identifier for each test result. The files should be spatially joined by intersection. The grid file should remain unchanged, and the case file should include the cell identifier, the test date, and the variant label.

Once the basic dataset is prepared, one last transformation should be undertaken. The case file must now be summarized by cell. To do this, the data should be group by cell and date, with the total count of tests and a count of each variant present. The resulting cell-case file will be used to perform the analysis.

 doi:10.20944/preprints202208.0209.v1

2.2. Analysis

The analysis subprocess can now be used to identify incidence rates and dominant viral variants. This paper describes how to create a dataset and perform basic analysis. However, more advanced spatial methods may be performed such as a Moran's I or Poisson regression. The analyst should consider if additional analyses would be useful and setup the workflow accordingly.

To complete the analysis, only the cell-case file will be used: the spatial grid file will be used in the render subprocess.

Using the cell-case file, a temporally smoothed incidence rate can be calculated. Smoothing is generally preferred since it reduces the effect noise may have on the analysis. However, if the sequencing of test represents a large portion of the total tests administered than smoothing may not be needed. The analyst should set a lookback parameter. This lookback parameter will be used for calculating a rolling average case count for each cell. Additionally, a rolling average case count should be calculated for each variant.

Using the rolling average, an incidence rate for each variant on each cell can be calculated. This is done by dividing the rolling average for the variant by the rolling average for the total population. If the rolling average of the total is zero, this will result in a division by zero error. If, therefore, the rolling average of the total case count is equal to zero, the incidence rate should be set to a null value to avoid division by zero errors.

Next, the dominant variant needs to be identified. The definition used is whatever variant has the highest incidence rate. In the case of a tie, the cell is coded as "No Dominant Variant." Using the incidence rate allows for greater flexibility with the definition of what cell is dominant. For instance, the definition can easily be coded as incidence rate greater than or equal to any percentage or that the difference between the highest incidence and the second must be greater than or equal to a set amount. This flexibility allows the analyst to configure the process to fit whatever circumstance they are facing.

After completing the analysis, a tabular file should be exported, with each row containing the cell identifier, the total count, the total incidence rate, the count and incidence rate for each variant, and the date.

2.3. Rendering

This subprocess is primarily concerned with rendering the data into the final presentation for interpretation. In this subprocess the spatial file and the cell-case file get joined and displayed. Joining should only happen after a date has been selected. Joining prior to date selection will result in a large computational burden which is not suitable for most standard devices. This is the rationale behind having separate files since a single file will, by necessity, contain a geography for every cell for every date.

While there are multiple methods of visualization, the two main visuals produced are a dominant variant map and a variant incidence rate map. The dominant variant map displays each cell with a sequenced test (from the rolling average total column) over an area base map and varies the color by variant label. This gives a visual representation of what variants are dominant in which areas. The incidence rate map is similar but with one difference: instead of displaying color by variant type, one variant is chosen and the cell color is varied by incidence rate based on a color ramp. Both these maps can be rendered as a static map by specifying a date. Usually, this will be the current date and will represent the newest available data. These can also be setup to be presented as an animation by having the presentation software automatically change the date. This animation helps in spotting how and where trends developed.

The author has rendered these maps in both ArcGIS Pro (v. 2.9) and Tableau (v. 2021.1). Both applications can be useful tools in rendering these maps. Tableau has the additional functionality of setting up "Pages" which provide for the visual animation. The "Pages" functionality allows for a time series to be created in which the disease progression is animated over the current map making it easy to see changes overtime using this functionality. Whether using tableau or another software visualization process, the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 August 2022 doi:10.20944/preprints202208.0209.v1

time series analysis maybe an important tool in understanding the disease development and progress throughout the study area.

3. Empirical Illustration: COVID-19 Sequencing at University Hospitals

On March 9th, 2020 the first case of COVID-19 was reported in Ohio(DeWine 2020). Rapidly, the virus spread throughout the community. University Hospital System of Cleveland (University Hospitals) was among the leaders coordinating the surveillance in Northeast Ohio. Started in 1866, University Hospitals is a research hospital affiliated with Case Western Reserve University which has served the Northeast Ohio community with over 150 locations which include hospitals, outpatient facilities, and primary care offices.

Early into the pandemic, it became clear that sequencing at a large scale would be necessary to best respond to ongoing emergency. University hospitals started sequencing tests on January 31st, 2021, averaging around 10% of all positive tests gathered being sequence to determine the specific coronavirus variant. This data was compiled daily and included the date of the test, the patient medical record number, and the World Health Organization COVID-19 variant label (e.g., Delta), among other data points. Using the medical record number, the patient residential location was added to this dataset. This provided the bases for the dataset that was needed to complete the visualization.

Using the software Alteryx Designer (v. 2021.1) the spatial grid was created for the University Hospitals market, which covers eighteen counties in Northeast Ohio. A shapefile of the University Hospitals market was used to define the spatial extent of the grid. A grid cell size of 1.6 km2 was used as this provided detail while still returning results that had enough sequenced tests to be considered useful. A smoothing period of seven days was used to reduce noise.

This data was then displayed inside Tableau (v. 2021.1). The spatial grid was first imported as a dataset and the geography information was added to a new worksheet and the cell identification was used in the detail plane to differentiate each cell. The incidence dataset was then imported and related to the spatial grid. The dominant variant type was used to determine the cell color with each variant given a unique color. Using the Tableau "Pages" functionality, the map was animated. The final product allows individual end users to examine the spatiotemporal history of COVID-19 variants over the Northeast Ohio area.

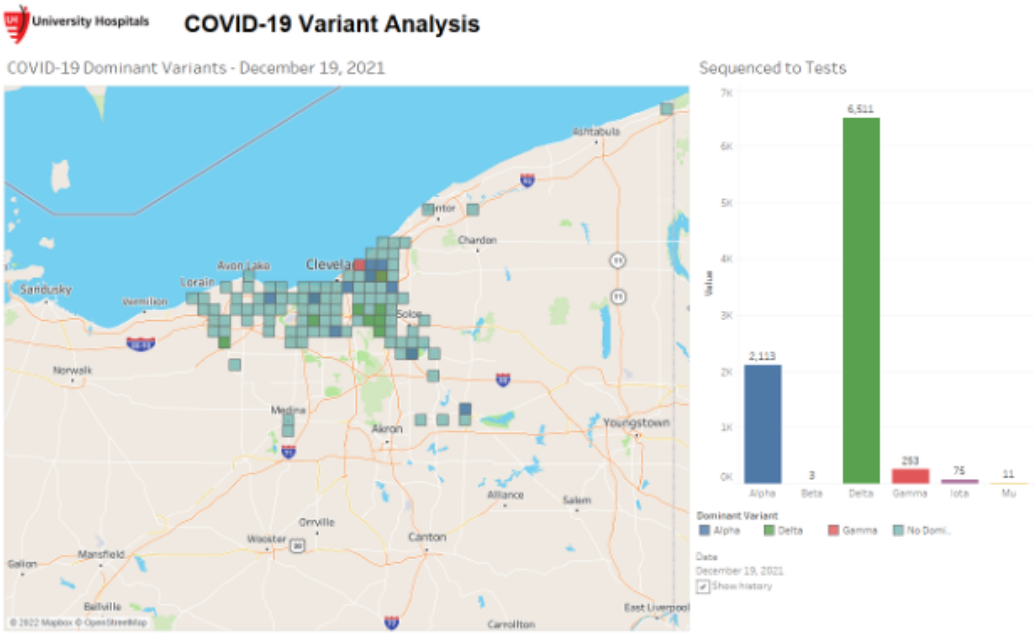

Figure 2. a example of the dashboard created at University Hospitals using simulated data

 doi:10.20944/preprints202208.0209.v1

Ultimately the data latency at University Hospitals limited this product to historical review only. This is an important point, this tool is only as good as the data it is provided. However, it was concluded that this tool may be useful in the future of monitoring other infectious diseases, such as influenza, which have a better data latency at University Hospitals.

4. Discussion

Viral variant analysis will continue to be a major aspect of disease surveillance. The method which is suggested above can help operationalize that analysis by providing a real-time data tool to track the variant's progression. Real-time surveillance is important to crafting response to disease development and allocating resources. Additionally, this surveillance will all decision leaders to construct narratives about the spread which will inform the intervention strategies used to arrest the development of the disease.

While the importance of disease surveillance has been recognized, there is little research into the methods of crafting a real-time surveillance reporting tool. In this paper such a tool was presented. The method suggested by the author gives a reasonably efficient way of examining and reporting on disease development in a specific area. This method can be deployed using publicly available tools, such as Python, or commonly available tools to most public health institutions, such as Tableau.

This tool was developed at University Hospitals during the SARS-CoV-2 outbreak. It was used to survey the developing variant evolution in gain actionable insights into how to best stage interventions.

This tool assumes that the data are readily available. this can be measured in the data latency which is the time between when a datum is available and when the report is made. The lower that time the faster and more accurate the data is to the current environment. well this tool cannot speed up institutional processes. It makes the reporting available as soon as the data is ready. This is because it automates all analysis tasks removing any need for an analyst to prepare this report prior to its consumption.

Any institution looking to implement this tool should consider carefully the environment in which they are instituting it. This is because parameters must be set according to the need of the moment. This includes the size of each grid cell and the lag effect overtime. As long as these parameters are set-up appropriately this report will provide useful information as soon as it is available.

This tool does have its limits. The most prevalent limit is data availability. That availability both encompasses the amount of data and the latency of the data (which was discussed above). While this tool is not meant to be Used for statistical analysis, care should be taken to make sure that the data gathered represents the underlying population. One way to do this is to compare the sequenced tests versus the total test administered versus the total population. this method allows for the analyst to make sure that the sequence tests represent not only the population but also the total number of tests administered. This also can be used to spot areas that are not receiving sufficient testing.

Gathering representative data may be particularly hard for smaller institutions without the resources to sequence a large number of tests. This tool may not be appropriate in situations where representative samples cannot be collected. However, adjusting the grid cell size may help representativeness at the expense of granularity. Additionally, hospital systems and public health institutions may band together to increase the effectiveness of their resource allocation. While there are barriers to sharing data, there may be ways of providing collaborative care without being non-compliant with regulations. Several institutions including University Hospitals participated in a data sharing agreement for SARS-CoV-2 testing.

As with any project that aggregates point data to a geography, in this case grid cells, the modifiable areal unit problem (MAUP) should be accounted for. The problem stems from the aggregation itself. one way to test to make sure that there is no MAUP is to change the geography by a fraction of a cell grid. if, for instance, your grid cell was one square

mile you could shift your geography by half a square mile in order to see if the results would be different. Should a drastically different map be generated after this adjustment you should examine your parameters to make sure that the MAUP is not a major factor in further analysis.

5. Conclusion

This data visualization tool and method allow decision makers the ability to get real-time data visualized in a fast-paced environment. This tool works best when the data are readily available, for infectious diseases, for institutions with the ability to provide interventions. This was the case during the height of the SARS-CoV-2 outbreak in Northeast Ohio when this tool was first built.

There are uses for this tool beyond SARS-CoV-2. This tool may be useful particularly for influenza surveillance. Since influenza is a fast spreading, seasonal disease, using this tool would allow the public health officials to quickly response to developing outbreaks. This could include ensuring the proper vaccines targeted at the specific variant of this disease are available, increase outreach and advocacy in communities experiencing high transmission rates, and quickly identifying areas where new variants are developing.

This tool, however, has limitations. The first limitation is data latency, which is the limitation that University Hospitals faced. Testing, sequencing, and reporting take time. In rapidly spreading diseases the time to get the data may be to long for effective real-time analysis. One core assumption this tool holds is that data is available in near-real-time. This limitation is less about the method and more about operational realities: this tool will not speed up an already complicated process. Before implementing this tool, any decision leaders should ensure that the processing time is sufficient to make sure that the reporting the dashboard provides is actionable.

A second limitation is the availability of data. Sequencing of viral variants at scale is a resource intensive task. Many smaller hospital systems and public health departments may lack the means to be able to sequence on the scale required to use this tool.

A third limitation is the modifiable areal unit problem (MAUP). This method is susceptible to the MAUP since the grid is completely arbitrary. In order to avoid the MAUP, multiple analyses with varied cell sizes should be conducted. Additionally, the extent can be extended by a fraction of a unit to vary the distinction between units. These analyses should be compared visually to ensure that similar spatial patterns are seen across the analyses.

While the environment is changing rapidly, surveillance must also be able to handle those rapid changes. This tool, if properly configured, can provide detailed real time analysis to the decision makers in a rapidly changing environment. additionally, this tool can be configured and deployed at most institutions at little to no additional infrastructure needed.

Acknowledgments: In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

Conflicts of Interest: The author declare no conflict of interest.

References

1. World Health Organization. WHO Timeline - COVID-19, 2020.
2. Baker, M. When Did the Coronavirus Arrive in the U.S.? Here's a Review of the Evidence. - The New York Times, 2020.
3. Centers for Disease Control and Prevention. SARS-CoV-2 Variant Classifications and Definitions, 2022.
4. Bioinformatics Resource Center. Virus Pathogen Database and Analysis Resource (ViPR) - Genome database with visualization and analysis tools, 2020.
5. Brister, J.R.; Ako-Adjei, D.; Bao, Y.; Blinkova, O. NCBI viral Genomes resource. *Nucleic Acids Research* **2015**, *43*, D571–D577. https://doi.org/10.1093/NAR/GKU1207.
6. McBurney, S.P.; Ross, T.M. Viral sequence diversity: challenges for AIDS vaccine designs. *Expert review of vaccines* **2008**, *7*, 1405. https://doi.org/10.1586/14760584.7.9.1405.

 doi:10.20944/preprints202208.0209.v1

7. Feuer, R.; Boone, J.D.; Netski, D.; Morzunov, S.P.; Jeor, S.C.S. Temporal and Spatial Analysis of Sin Nombre Virus Quasispecies in Naturally Infected Rodents. *Journal of Virology* **1999**, *73*, 9544. https://doi.org/10.1128/JVI.73.11.9544-9554.1999.
8. Carroll, M.W.; Matthews, D.A.; Hiscox, J.A.; Elmore, M.J.; Pollakis, G.; Rambaut, A.; Hewson, R.; García-Dorival, I.; Bore, J.A.; Koundouno, R.; et al. Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. *Nature 2015 524:7563* **2015**, *524*, 97–101. https://doi.org/10.1038/nature14594.
9. Dellicour, S.; Rose, R.; Faria, N.R.; Vieira, L.F.P.; Bourhy, H.; Gilbert, M.; Lemey, P.; Pybus, O.G. Using Viral Gene Sequences to Compare and Explain the Heterogeneous Spatial Dynamics of Virus Epidemics. *Molecular Biology and Evolution* **2017**, *34*, 2563–2571. https://doi.org/10.1093/MOLBEV/MSX176.