Spatial Data In the Age of COVID-19. Lessons learned in managing, serving, and teaching with public health data.

Stacy Curry-Johnson, PhD.[a], Natalie Robbins, PSM[b] and Steven A. Wernke, PhD[c]

[a] Library, Vanderbilt University, Nashville, TN, US stacy.e.curry@vanderbilt.edu *
[b] Spatial Analysis Research Lab (SARL), Vanderbilt University, Nashville, TN, US*
[c] Department of Anthropology, Vanderbilt University, Nashville, TN, US

Keywords: Data Dashboard; GIS; COVID-19; Geovisualization; Open Data; Public Health; Geospatial

Introduction

As contagious disease transmission is an inherently spatial process, the COVID-19 pandemic expanded the demand for spatial datasets and visualizations, for public education, epidemiological knowledge, and policy guidance. Researchers, policymakers, and the public have requested quick and accurate information and easy access visualisation. Incorporating the spatial dimensions of the data aids in public understanding, dynamically and responsively. However, as data are disseminated and archived by multifarious entities, issues of data trust, data formatting and management can be challenging.

In recent years, especially with the roll out of the ArcGIS Hub technology, open data sites have become commonplace for local, state, and federal data repositories. By hosting data on these sites, data can be made available in geospatial formats (shapefiles, feature classes), tabular formats compatible with statistical software systems (.csv, .txt), and even extensional spatial formats such as kml/kmz files for Google Earth. In the digital age where information is rapidly changing, the ability to access authoritative geospatial data quickly and easily is critical for communicating.

In the very dynamic and fraught early days of the pandemic, researchers reached for tools to integrate disparate data streams. One of the most ubiquitous solutions has been cloud-based GIS "dashboard" platforms. The urgency of the need to visualise rapidly changing data spatially superseded the quality and ethical sourcing checks that are the standard. With multiple sources of data comes potential propagation in error in not only data collection but formatting and dissemination. Data of various formats, sources, quality, and accuracy must be "wrangled" or standardised to work within the parameters of the ArcGIS Dashboard technology.

Compounding the data quality challenges is the ethical representation of variables. Along with the need to bring together not only dynamically changing case counts, but requests for spatial analysis of other variables. Such variables include healthcare available, insurance, demographics, socio-economic factors, etc. Critical to these enhancements is the maintaining of the ethical representation minimizing erroneous correlation.

This paper presents reflections and recommendations for spatial data management and presentation based on our recent experience in the rapid development and deployment of a COVID-19 dashboard focusing on the State of Tennessee (https://tn-covid19-vanderbilt.hub.arcgis.com/) since the early days of the pandemic.

Method

In Spring 2020 as the pandemic evolved and a two-week nationwide quarantine was announced, our team recognized the need for an up to date, county-wide map showing cases across the state of Tennessee. At this time, there had yet to be an official case spread map put out by the state of Tennessee. Drawing inspiration from the global Johns-Hopkins COVID-19 dashboard (John Hopkins University & Medicine, 2022), the Vanderbilt Initiative for Interdisciplinary Geospatial Research (VIIGR) began pulling daily case count data from the Tennessee Department of Health (TDH) website and spatializing it. TDH disseminated data at the county-level, originally as a PDF, later being made available as an Excel (.xlsx) file. New case count data was copied from TDH data, transposed, and appended to two hosted CSV tables on the Tennessee COVID-19 Data Portal. A separate hosted feature layer was appended with daily case counts, deaths, negative test results and inactive or recovered cases.

Along with displaying county-level case counts, the dashboard incorporated Census data of county demographic and socioeconomic factors, including percentage of population uninsured, percentage of population over the age of 65, and percentage of population receiving SNAP benefits. Demographic and socioeconomic data was pulled from the American Community Survey (ACS) 2018 estimates and Tennessee Division of TennCare excel files and normalized as a function of total county population. These metrics are useful for understanding the effect of COVID-19 in reference to risk factors and proximity to health care facilities.

The dashboard also incorporated super-spreader, group facilities, such as prisons and assisted living facilities. Case data for prisons was updated based-off Tennessee Department of Corrections (TDOC) weekly PDF reports. Assisted living facilities outbreak data was pulled from TDH as an HTML table. Locations were geocoded using ArcGIS Pro and appended to the Tennessee Department of Corrections Facility COVID-19 Outbreaks or Tennessee Assisted Living Facility hosted feature layers.

Results

Community spread of COVID-19 in Tennessee began on March 5, 2020. In the weeks following, TDH began publishing county-level data on COVID-19 case counts. On March 16, 2020, VIIGR stood up a Tennessee-specific, ArcGIS Dashboard visualising the case counts of COVID-19 in Tennessee. The Tennessee COVID-19 dashboard contained two web maps: confirmed COVID-19 cases and active COVID-19 cases in Tennessee. From this grew the Tennessee COVID-19 Data Portal, to house COVID-related open data sources and visualizations.

By April, the State of Tennessee had stood up their own COVID-19 dashboard. Although this dashboard was fed by TDH case data, no hosted feature layers (REST APIs) were publicly disseminated. Throughout the pandemic, data was only made

available through the TDH site as an Excel file— a data format that is not easily ingestible by geographic information systems or statistical software platforms. Hosted feature layers created by VIIGR for this project reached over 1 million requests at the height of the pandemic, highlighting the need for accessible, GIS-ready data.

Discussion and Conclusion

Being both a user and creator of COVID-19 data highlighted the need for a discussion around the dissemination of public health data. Throughout the maintenance of the TN COVID-19 portal, the frequency of data updates and types of data being disseminated fluctuated greatly. The constantly evolving nature of the COVID-19 crisis presented unique challenges for data users. As new metrics were emphasised, the data reported by TDH changed to reflect that. For example, hospitalization rates became a focus among public health experts, TDH integrated that metric into their public dataset (TN Department of Health, 2022).

Additionally, the format data was disseminated from TDH evolved from initially being presented as an HTML table, to a PDF report, and finally as an excel file, causing us to have to adapt the way this data is accessed and integrated. The final excel (.xlsx) format presents challenges, as it cannot be directly ingested into GIS desktop or cloud platforms. This introduces an additional step in data processing to convert from .xlsx to .csv format.

In Tennessee, COVID-19 data is also reported differently by state and local agencies, contributing to confusion among data consumers. Nashville-Davidson County COVID-19 counts are a prime example of the difference between state and local reporting. Nashville Metro Health reports daily COVID-19 counts as a total of anyone who is currently in Davidson County and has tested positive for COVID-19. The Tennessee Department of Health breaks down cases and separates out COVID-19 cases into Davidson County residents and out of state cases. Therefore, someone who is a visitor to the city, or has no permanent Davidson County address would be counted in the Metro Nashville Davidson case counts but would not be counted in the TDH Davidson case count. If data users were to compare the Metro Nashville dataset and TDH dataset, they would see smaller Davidson COVID-19 case counts in the TDH dataset. Further compounding potential confusion, may be lags in data reporting between the local health departments and TDH, causing data presented in the TDH dataset to be delayed several days.

When a public health crisis, such as the COVID-19 pandemic, strikes, quick response and dissemination of information is key. It is because of this that we recommend the adoption of standards for the dissemination of data. Our recommendations are as follows:

1. Data publishers disseminate case counts and associated attribute information as a CSV file that can be easily accessed by geospatial and statistical analysts.
2. Information can be served out through an Open Data Portal maintained by the state. Most states already have these open data portal sites configured and some are already serving out COVID-19 data on these platforms, with the state of New York being an example (cite).

3. Standardize how reporting is done across the state (or nationally through CDC guidelines), to ensure state and local health department numbers match. This will increase public trust in the data being disseminated and eliminate the confusion about numbers that do not match up on a local and state scale.

As dashboards have taken off in every state it is important to assess the nature of how they are created and what data standards exist to support them. A future contribution would be to conduct a review of COVID-19 dashboards that were stood up across the US and identify strengths and challenges, while further developing strategies for standards moving forward with pandemic and public health geovisualization shed light on how to improve data accuracy and geovisualization for future health crises.

References

Covid-19 Dashboard. Johns Hopkins Coronavirus Resource Center. (n.d.). Retrieved June 20, 2022, from https://coronavirus.jhu.edu/map.html

Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis. 20(5):533-534. doi: 10.1016/S1473-3099(20)30120-1

Lv, Y., Ma, C., Li, X., & Wu, M. (2021). Big data driven COVID-19 pandemic crisis management: potential approach for global health. *Archives of medical science : AMS*, *17*(3), 829–837. https://doi.org/10.5114/aoms/133522

"TDH Announces First Case of COVID-19 in Tennessee." Accessed June 20, 2022. https://www.tn.gov/health/news/2020/3/5/tdh-announces-first-case-of-covid-19-in-tennessee.html.

U.S. Census Bureau. (2020). 2018 American Community Survey 5-year Estimates [CSV Data file]. Retrieved from https://data2.nhgis.org/downloads.

Z. -H. Zhou, N. V. Chawla, Y. Jin and G. J. Williams, "Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives [Discussion Forum]," in IEEE Computational Intelligence Magazine, vol. 9, no. 4, pp. 62-74, Nov. 2014, doi: 10.1109/MCI.2014.2350953.