HUD Crosswalk Files Facilitate Multi-State Census Tract COVID-19 Spatial Analysis

Alexander Din
U.S. Department of Housing and Urban Development

Ron Wilson
University of Maryland, Baltimore County

The views expressed in this article are those of the author and do not represent the official positions or policies of the Office of Policy Development and Research, the U.S. Department of Housing and Urban Development, or the U.S. Government.

Abstract

The coronavirus COVID-19 has infected millions of Americans. Datasets like the national county-level aggregation of COVID-19 case counts that Johns Hopkins University & Medicine assembled have been widely used, but few analyses have been performed at the local level due to the low supply of data. Like many things American, the distribution of COVID-19 data varies due to differing state, county, and local government reporting policies. The result is a patchwork of COVID-19 data at the local level, mostly aggregated to ZIP Codes due to ease of data processing rather than census tracts which are a better geographical unit for analysis. Local level COVID-19 data are rare and often only available for small areas. In this article, we demonstrate how the U.S. Department of Housing and Urban Development (HUD) Crosswalk Files can be used to assemble a census tract-level dataset of COVID-19 case rates in the Washington, D.C. Metropolitan Statistical Area across multiple states.

Coronavirus Data

The most common COVID-19 dataset used for geospatial analysis has been the county-level aggregation of COVID-19 cases that Johns Hopkins University assembled.[1] This dataset has national coverage, but the observations are counties, which are not granular and vary greatly in

[1] https://github.com/CSSEGISandData/COVID-19

size, shape, and demographics. For the few states and local governments that have released local-level COVID-19 data, most datasets are compiled at the ZIP Code-level.[2] Data aggregations to ZIP Codes are common because ZIP Codes are commonly recorded with patient record files, and tabulation at these geographies, which frequently contain thousands of households, helps to preserve privacy. In contrast, determining the census tract for a patient requires geocoding patient addresses, a process that requires a sophisticated geographic information system, technical staff, and operating costs.

Much local-level COVID-19 spatial analysis has focused on ZIP Code analysis of COVID-19 cases in New York City (NYC). While NYC was experiencing the first outbreak in the United States, the NYC Health Department began providing COVID-19 data to the public.[3] This release of data led to a number of studies focusing on NYC, suggesting the NYC subway spread the virus (Din and Wilson, 2020a; Harris, 2020); indicating that neighborhoods with greater rates of certain occupations experienced greater rates of COVID-19 cases (Almagro and Orane-Hutchinson, 2020); per capita income is negatively correlated with COVID-19 case rates (Olmo and Sanso-Navarro, 2020); and patients living in poorer neighborhoods or areas with a greater Black or immigrant population were more likely to test positive but less likely to get tested (Borjas, 2020). A search of Google Scholar for "zip code coronavirus" from 2020 onward will yield results mostly discussing NYC.

Local-level analysis in other jurisdictions are few and far between. In Milwaukee, COVID-19 case counts were greater in predominantly Black neighborhoods (Rast, 2020). In Texas, poverty rates were strongly correlated with COVID-19 cases in Bexar County/San Antonio, whereas workers using public transportation were highly correlated in Harris County and Fort Bend County, and socially vulnerable populations were positively correlated across all jurisdictions (Chen and Jiao, 2020).

Washington, D.C. Metropolitan Statistical Area

Two commonalities happen among many local-level spatial analyses of COVID-19. First, analyses typically use ZIP Codes because they are convenient for data aggregation even though they are frequently inadequate for spatial analysis (Beyer, Schultz, and Rushton, 2007; Cudnick et al., 2012; Grubesic and Matisziw, 2006; Krieger et al., 2002; Oregon Health Authority, 2020; Sadler, 2019; Wilson, 2015) Second, local-level analyses focus on few areas, mostly NYC. Although it is difficult to get publicly available COVID-19 data in many jurisdictions, we demonstrate, as a new example, that such data are available across the vast majority of the multi-state Washington, D.C. Metropolitan Statistical Area (MSA) and its component jurisdictions in the District of Columbia, Maryland, and Virginia. We also demonstrate that it is possible to adequately estimate such data at the census-tract level by cross-walking the ZIP Code counts to census tracts to avoid the geographic problems that occur with ZIP Codes (see Din and Wilson, 2020b; Wilson and Din, 2018, for more on crosswalking ZIP Code data).

[2] It is worth noting that Wisconsin provides COVID-19 data at the census-tract level and perhaps could offer technical assistance to other states and jurisdictions for how to aggregate and distribute census tract data. https://data.dhsgis.wi.gov/datasets/40a25761793c4501a291852b7d39432b_9

[3] https://github.com/nychealth/coronavirus-data

The Washington, D.C. MSA is unique because it is centered around the District of Columbia, a federal district that is a city that operates like a county and state but is legally neither, and the bulk of the region's population is outside of the city in suburban Maryland and Virginia, and a small portion in West Virginia. The state-equivalents must frequently work together on issues that affect the region due to their high level of interconnection, but such cooperation is often difficult due to differing data standards and policies. During the COVID-19 pandemic, as the nation shut down, each jurisdiction enacted and enforced restrictions and procedures separately from each other. To complicate matters further, the state governments of Maryland and Virginia have allowed counties autonomy to remain in stricter lockdown procedures as the counties saw fit.

Data were collected cumulatively through October 1, 2020, from Maryland's iMap Open Data Portal (MD iMap), the Virginia Open Data Portal, and the District of Columbia's Coronavirus Dashboard, although the official first-reported COVID-19 cases varied across jurisdictions. Local-level data were unavailable for Jefferson County, West Virginia. Data from each jurisdiction were available in different formats.

Maryland offers multiple COVID-19 related datasets on MD iMap. COVID-19 case-count data are available as a cumulative daily count aggregated to ZIP Codes and are available via a modern, easily accessible Esri data portal.[4]

In Virginia, COVID-19 data are offered regarding positive COVID-19 cases, and COVID-19 testing encounters aggregated to ZIP Codes in a single dataset.[5] Data from Virginia included daily cumulative cases across the reporting time period and were available via multiple methods from a Socrata open data portal.

Data for the District of Columbia differed in multiple ways because they were available via a tabular download from the District's Coronavirus Dashboard,[6] providing only cumulative counts for the current day, and the data were aggregated to neighborhoods instead of ZIP Codes.

ZIP Code data for Maryland and Virginia were crosswalked to census tracts from ZIP Codes using the U.S. Department of Housing and Urban Development (HUD) U.S. Postal Service (USPS) Crosswalk Files,[7] a reasonable method for estimating data at the census-tract geography from ZIP Code geographies (Din and Wilson, 2020a). In the District, because neighborhoods are aggregations of census tracts, data were crosswalked to census tracts using proportional ratios of population between the neighborhood and its component census tracts using 2014–2018 American Community Survey (ACS) 5-Year Estimate data.

Results

Exhibit 1 and exhibit 2 map the rate of COVID-19 cases per 10,000 residents in census tracts across the Washington, D.C. area. Exhibit 1 shows higher rates of COVID-19 cases closer to and within the District, although there are pockets of higher case rates in northern and eastern

[4] https://data.imap.maryland.gov/datasets/mdcovid19-master-zip-code-cases/data

[5] https://data.virginia.gov/Government/VDH-COVID-19-PublicUseDataset-ZIPCode/8bkr-zfqv/data

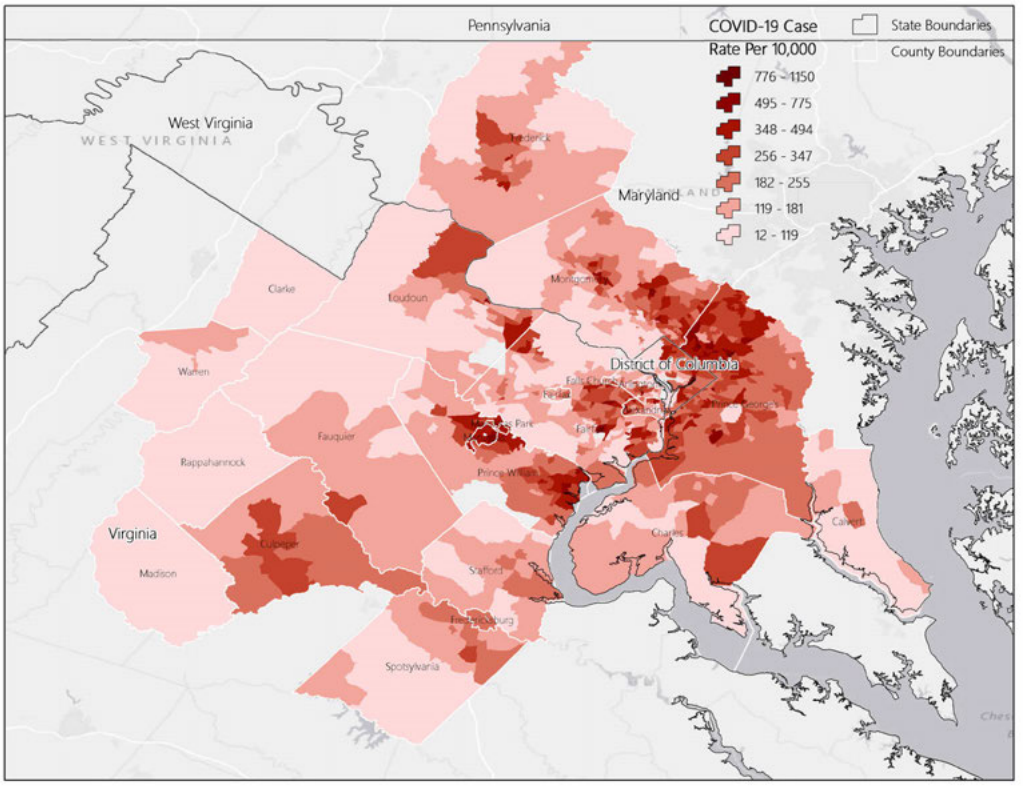
[6] https://coronavirus.dc.gov/data

[7] https://www.huduser.gov/portal/datasets/usps_crosswalk.html

Montgomery County, wide swaths of Prince George's County, eastern Fairfax County, Manassas and Manassas Park, and eastern Prince William County.

Exhibit 1

COVID-19 Case Rate Per 10,000 Population in the Washington, D.C. Metropolitan Statistical Area.

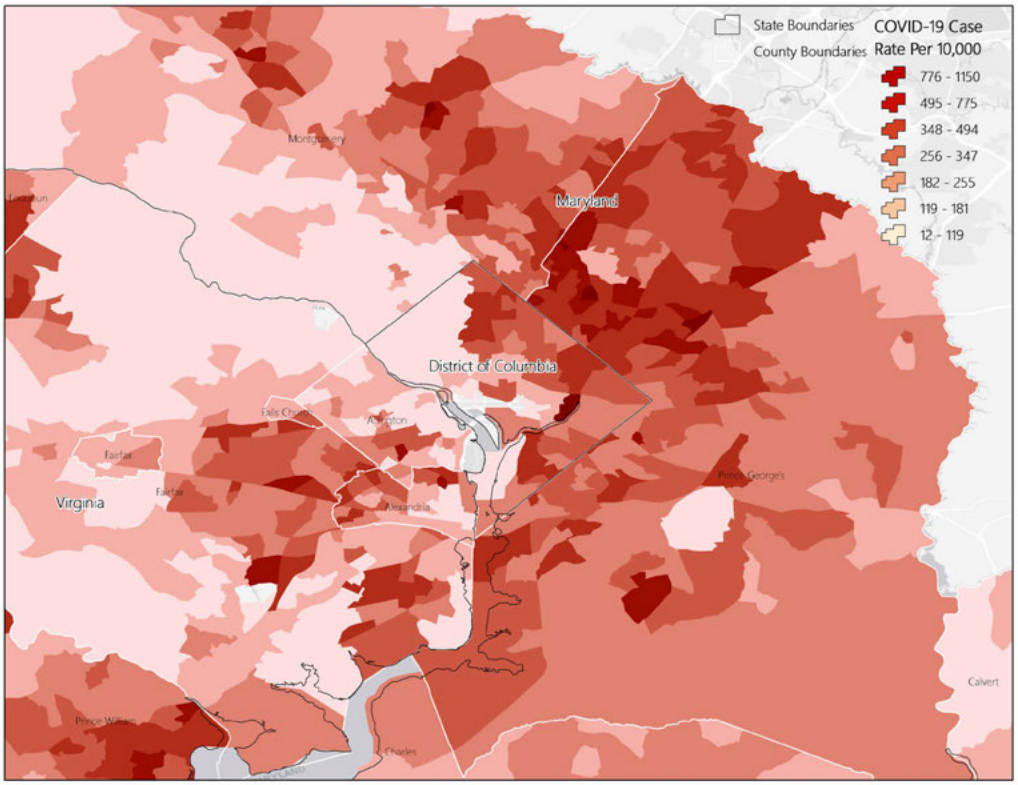


Sources: COVID case-rate data – District of Columbia Coronavirus Dashboard; Maryland iMap; Virginia Open Data Portal

Exhibit 2, which focuses on Washington, D.C., shows that, although much of the District has higher COVID-19 case rates, large swaths of census tracts in neighboring suburban counties have similar or greater case rates. In particular, northern Prince George's County has many census tracts that exceed the rate in the center of the metropolitan area. This area has been the regional center for many immigrant communities spanning several decades (Price et al., 2005).

Exhibit 2

COVID-19 Case Rate Per 10,000 Population in the Washington, D.C. and Nearby Suburbs.



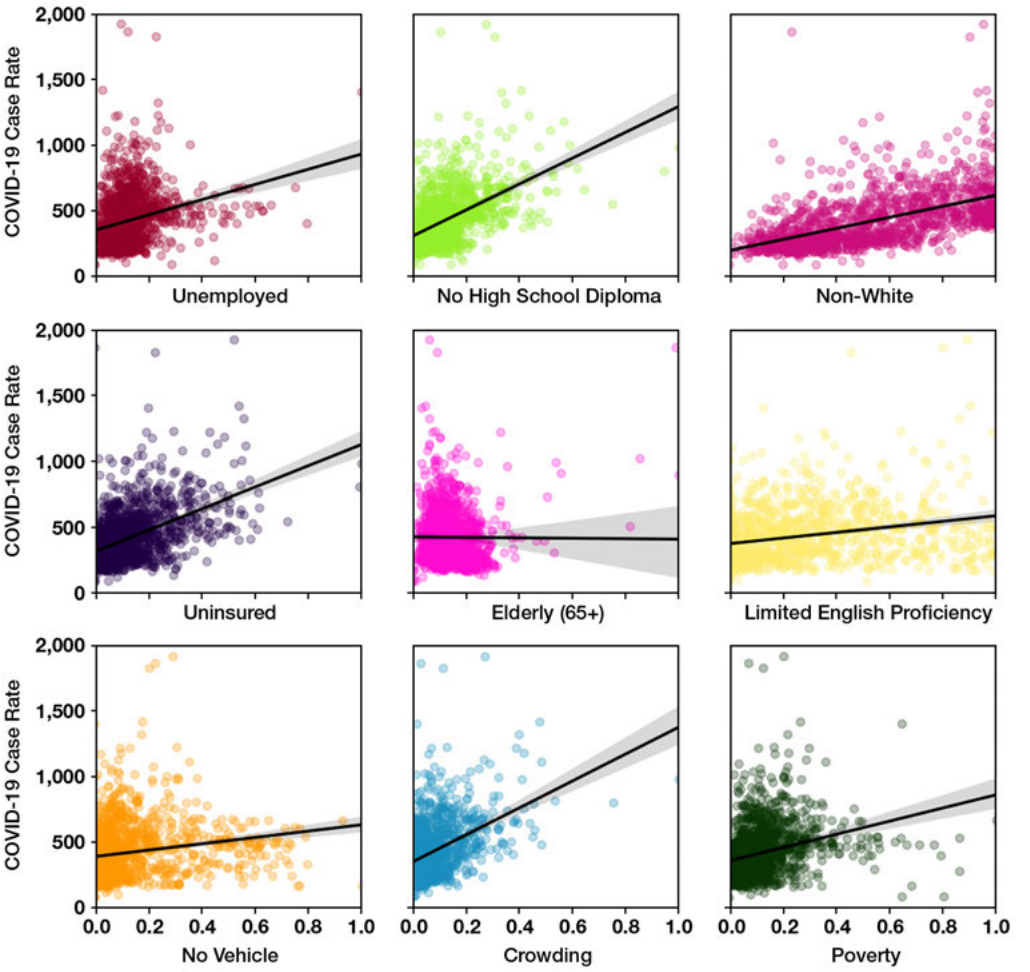
Sources: COVID case-rate data – District of Columbia Coronavirus Dashboard; Maryland iMap; Virginia Open Data Portal

Exhibit 3 is a set of regression plots comparing the COVID-19 case rate per 10,000 population to component variables in the Centers for Disease Control and Prevention Social Vulnerability Index[8] (SVI). Because the COVID-19 case rates have been estimated at the census-tract level, linking the SVI is a simple task because it is produced at the census-tract level. SVI variables were commonly used in analyses with COVID-19 across research articles and studies. Many of the SVI variables, particularly the rate of those without a high school diploma, those who lack medical insurance, and households with more members than bedrooms, correlate strongly with COVID-19 case rates. The rate of people aged 65 or older did not correlate strongly with COVID-19 case rates, but this may be due to the median age of COVID-19 patients declining as the pandemic progresses (Boehmer et al., 2020).

[8] https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html

Exhibit 3

Social Vulnerability Index Variables and COVID-19 Case Rate Per 10,000 Population.



Sources: Variables and COVID case-rate data – District of Columbia Coronavirus Dashboard; Maryland iMap; Virginia Open Data Portal; Centers for Disease Control

Summary

Our analysis shows that it is possible to estimate COVID-19 case rates without relying on the use of ZIP Codes. The results show much more detailed and robust map patterns to assess the distribution of infection rates across the region. The use of the estimates at the census-tract level also now allows for analyses with other data to explore the connections between infection rates and demographics.

Notes

The authors did not summarize all spatial research related to COVID-19 but merely intended to provide an overview of local level spatial research conducted.

Information for the Social Vulnerability Index rate variables is available at: https://svi.cdc.gov/Documents/Data/2018_SVI_Data/SVI2018Documentation-508.pdf

Acknowledgments

The authors would like to thank those who make scholarly articles publicly available.

Authors

Alexander Din is a social science analyst in the Office of Policy Development and Research at the U.S. Department of Housing and Urban Development.

Ron Wilson is an adjunct faculty member and acting director of the Geographic Information Systems Program at the University of Maryland, Baltimore County.

References

Almagro, Milena, and Angelo Orane-Hutchinson. 2020. "The Determinants of the Differential Exposure to COVID-19 in New York City and Their Evolution Over Time." Social Science Research Network. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3573619.

Beyer, Kirsten, Alan Schultz, and Gerard Rushton. 2007. "Using ZIP Codes as Geocodes in Cancer Research." In *Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research and Practice*, edited by G. Rushton, M. P. Armstrong, J. Gittler, B. R. Greene, C. E. Pavlik, M. M. West, and D. L. Zimmerman. Boca Raton, FL: CRC PRESS: 37–64.

Boehmer, Teagan .K., J. DeVies, E. Caruso et al. 2020. "Changing Age Distribution of the COVID-19 Pandemic—United States, May–August 2020," *MMWR Morb Mortal Wkly Rep* 2020 69:1404–1409. DOI: http://dx.doi.org/10.15585/mmwr.mm6939e1.

Borjas, George. 2020. Demographic Determinants of Testing Incidence and COVID-19 Infections in New York City Neighborhoods. Working paper. National Bureau of Economic Research. https://www.nber.org/papers/w26952.pdf.

Chen, Yefu, and Junfeng Jiao. 2020. "Relationship Between Socio-Demographics and COVID-19: A Case Study in Three Texas Regions," Social Science Research Network. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3636484.

Cudnik, Michael T., Jing Yao, Dana Zive, Craig Newgard, and Alan T. Murray. 2012. "Surrogate Markers of Transport Distance for Out-of-Hospital Cardiac Arrest Patients," *Prehospital Emergency Care* 16 (2): 266–272.

Din, Alexander, and Ron Wilson. 2020a. "Using HUD Crosswalk Files to Improve COVID-19 Analysis at the ZIP Code and Local Level," *Cityscape: A Journal of Policy Development and Research* 22 (3): 365–371.

———. 2020b. "Crosswalking ZIP Codes to Census Geographies: Geoprocessing the U.S. Department of Housing & Urban Development's ZIP Code Crosswalk Files," *Cityscape: A Journal of Policy Development and Research* 22 (1): 293–314.

Grubesic, Tony H. and Timothy C. Matisziw. 2006. "On the Use of ZIP Codes and ZIP Code Tabulation Areas (ZCTAs) for the Spatial Analysis of Epidemiological Data," *International Journal of Health Geographics* 5 (1): 58.

Harris, Jeffrey. 2020. The Subways Seeded the Massive Coronavirus Epidemic in New York City. Working paper. National Bureau of Economic Research. https://www.nber.org/papers/w27021.pdf.

Krieger, Nancy, Pamela Waterman, Jarvis T. Chen, Mah-Jabeen Soobader, S.V. Subramanian, and Rosa Carson. 2002. "Zip Code Caveat: Bias Due to Spatiotemporal Mismatches Between Zip Codes and U.S. Census-Defined Geographic Areas—The Public Health Disparities Geocoding Project," *American Journal of Public Health* 92 (7): 1100–1102.

Olmo, Jose and Marcos Sanso-Navarro. 2020. "Modelling the spread of COVID-19 in New York City" Social Science Research Network. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3713720.

Oregon Health Authority. 2020. COVID-19 Weekly Report, May 5, 2020. https://www.oregon.gov/oha/PH/DISEASESCONDITIONS/DISEASESAZ/Emerging%20Respitory%20Infections/COVID-19-Weekly-Report-2020-05-05-FINAL.pdf.

Price, Marie, Ivan Cheung, Samantha Friedman, and Audrey Singer. 2005. "The World Settles In: Washington, DC, as an Immigrant Gateway," *Urban Geography* 26 (1): 61–83.

Rast, Joel. 2020. "Milwaukee's Coronavirus Racial Divide." https://uwm.edu/ced/wp-content/uploads/sites/431/2020/04/COVID-report-final-version.pdf.

Sadler, Richard. 2019. "Misalignment Between ZIP Codes and Municipal Boundaries: A Problem for Public Health," *Cityscape, A Journal of Policy Development and Research* 21 (3): 335–340.

Wilson, Ronald E. 2015. "The Neighborhood Context of Foreclosures and Crime," *Cartography and Geographic Information Science* 4 (2): 162–177.

Wilson, Ron and Alexander Din. 2018. "Understanding and Enhancing the U.S. Department of Housing and Urban Development's ZIP Code Crosswalk Files" *Cityscape: A Journal of Policy Development and Research* 20 (2): 277–294.