RSPP Working Paper

Socio-Demographic Disparities in COVID-19 Case Rates and Testing: An exploratory spatial analysis of ZIP code data in Chicago, IL

Kevin CREDIT

Nº 2020.002 – Special Series on Covid-19

Regional Science **Policy&Practice**
Regional Science Associaton International

Socio-Demographic Disparities in COVID-19 Case Rates and Testing: An exploratory spatial analysis of ZIP code data in Chicago, IL

By Kevin Credit[1]

ABSTRACT

The recent release of ZIP code level data on the novel coronavirus (COVID-19) provides an important opportunity to look at neighborhood-level socio-demographic and economic correlated of case and testing rates. This exploratory spatial analysis finds that black neighborhoods and those with older residents and higher proportions of employment in service occupations tend to have higher case rates (even when controlling for access to testing sites), while Hispanic neighborhoods have a lower than expected level of testing for COVID-19, even when controlling for the level of observed infection activity.

KEYWORDS

COVID-19, social determinants of health, spatial access, spatial econometrics

1. Introduction and Background

The novel coronavirus SARS-CoV-2 (COVID-19) was first discovered in 2019 in the Hubei province of China and has since spread across the world, creating a global pandemic that has resulted in over 2,770,000 confirmed cases and over 190,000 deaths (as of April 24, 2020) (Worldometer, 2020). While significant research effort has been dedicated to tracking the dynamics of the spread of the virus at the national and county levels (JHU, 2020; Dong, Du, & Gardner, 2020; CSDS, 2020), data at smaller spatial scales – such as ZIP code level – has only recently been made available in many jurisdictions, allowing for neighborhood-level analyses. These data provide the first opportunity to begin to explore the spatial associations between socio-demographic and economic characteristics and neighborhood-level virus activity.

Neighborhood-level analyses are particularly important in the context of social determinants of health (SDOH), an emerging body of research that recognizes that the social and economic conditions in which people live shape their ultimate health outcomes in a number of critically-important ways (Kolak et al., 2020; Solar & Irwin, 2010). In the US, historic and continuing patterns of racial and ethnic segregation often overlap with environmental degradation and socioeconomic deprivation, creating a context in which non-white communities are often systematically faced with higher rates of chronic disease, lower levels of infrastructure investment, lower access to healthcare and healthy food options, higher crime, and lower life expectancies (Braveman et al., 2010; Williams et al., 2010; Williams, Priest, & Anderson, 2016).

As one of the most segregated regions in the US, the Chicago, IL, metropolitan area provides an important case in which to explore the possible socio-demographic disparities in COVID-19 case rates and testing: according to a recent analysis of life expectancy at the Census tract scale, life expectancies in Cook County (the central county in the region containing the City of Chicago) range from 59.9 (lower

[1] Kevin Credit, Assistant Instructional Professor in GIScience and Assistant Director for Urban Informatics, Center for Spatial Data Science, University of Chicago, 1155 E 60th St, Room 211A, Chicago, IL, USA (e-mail: kcredit@uchicago.edu).

than countries such as Angola, Mali, and Zimbabwe) to 90 (which, as a country, would be the highest in the world - higher than Monaco at 89.4) (Arias et al., 2018; NCHS, 2015; CIA Factbook, 2017).

Following this pattern, hospitalizations and deaths from COVID-19 at the individual case level have — to this point — been strongly associated with minority groups and those with underlying conditions. According to the Chicago Department of Public Health (CDPH), as of April 23, 91.1% of Chicago residents who have died from COVID-19 had evidence of at least one chronic underlying health condition, and 75.9% of deaths have been from Latinx or black residents, compared to only 17.7% of deaths from white residents (2020).

Given these initial indications of disparate outcomes – and the fact that those in minority-majority neighborhoods may be more likely to work in service occupations for essential businesses that are not able to be shut down during Illinois' ongoing state-wide stay-at-home order – this paper explores the spatial relationships between black and Hispanic neighborhoods, case rates, testing rates, and other relevant age-related, transportation, and economic features in order to better understand the current impact of COVID-19 in the Chicago region and its burden on particular neighborhoods. Indeed, the exploratory results suggest that black neighborhoods and those with older residents and higher proportions of employment in service occupations tend to have higher case rates (even when controlling for access to testing sites), while Hispanic neighborhoods have a lower than expected level of testing for COVID-19, even when controlling for the level of observed infection activity.

2. Data Preparation

The data for this paper come from a variety of sources: the number of COVID-19 tests and confirmed cases by ZIP code as of April 16th, 2020, come from the website of the Illinois Department of Public Health (IDPH, 2020). This information was joined to Census data at the ZIP code Tabulation Area (ZCTA) from the 2014-2018 American Community Survey (ACS) for ZIP codes within the Illinois portion of the Chicago Core-Based Statistical Area (CBSA), i.e., the greater Chicago region. A range of Census variables were originally considered, including: percent population aged 0 – 44, 44 – 64, and 65+; percent white (non-Hispanic), black (non-Hispanic), and Hispanic population; percent of workers commuting by various modes, including automobile, transit, walking/biking, and telecommuting (working from home); and the percent of workers in various occupations, including "office" and technical occupations, healthcare service, protective service (such as fire fighters and police officers), food preparation, cleaning service, and transportation. Data on the location of COVID-19 testing sites (as of March 16th) in the Chicago region were also obtained from an open-source workbook compiling the location of testing sites in the US, as there was at the time no official government documentation of testing sites available (COVID-19 Drive Thru Location Tracking, 2020).

Given the collinearity of several of these sets of predictor variables, Principle Components Analysis (PCA) was used in three cases to reduce the number of data dimensions and combine variation from several correlated variables into a smaller number of uncorrelated components. The loadings of the three PCA analyses using the Singular Value Decomposition (SVD) method in GeoDa v.1.14 with a Z-score transformation are shown in Table 1.

Components explaining a large amount of variation that matched conceptually with features of interest were retained; specifically, of the transportation variables, PC1 (PC_TPT) was positively related to transit and pedestrian commuting and teleworking, thus capturing the higher income urban professional population who use transit but are also able to work from home and afford to live close enough to work to walk. PC2 (PC_TRANS), on the other hand, related positively only to transit commuting, capturing a different type of (likely) lower income transit-dependent population. For the age variables, PC1 (PC_YOUNG) generally captures younger populations, while PC2 (PC_OLD) captures older populations;

transforming these variables using the PCA ensures that they are uncorrelated with one another and thus able to be used in a spatial econometric model. Finally, PC1 (PC_SRV) of the occupation variables was retained, as it captured the most important feature of the input occupational data: a negative association with technical occupations (most likely to be able to work remotely during the pandemic) and positive associations with each of the other specific service occupations, chosen because they are all related to essential services that cannot be performed at home during a pandemic.

TABLE 1. PCA Variable Loadings for Principle Components (components retained in bold)

	Original Variables	PC1	PC2	PC3	PC4	PC5	PC6
Transportation Variables	% Auto	**-0.637**	**-0.124**	0.011	0.761		
	% Transit	**0.457**	**0.646**	-0.362	0.493		
	% Pedestrian and Bicycling	**0.512**	**-0.241**	0.733	0.379		
	% Working from Home	**0.351**	**-0.714**	-0.576	0.187		
Age Variables	% 0-44	**0.656**	**-0.049**	0.753			
	% 45-64	**-0.510**	**-0.765**	0.393			
	% 65+	**-0.557**	**0.642**	0.527			
Occupation Variables	% Office and Technical	**-0.527**	0.044	-0.050	-0.119	-0.160	0.823
	% Healthcare Service	**0.383**	0.398	-0.368	-0.739	0.046	0.104
	% Protective Service	**0.220**	0.537	0.792	0.020	-0.127	0.139
	% Food Preparation	**0.414**	-0.498	0.084	-0.121	-0.735	0.136
	% Cleaning Service	**0.448**	-0.413	0.179	0.022	0.631	0.445
	% Transportation	**0.394**	0.365	-0.442	0.651	-0.132	0.274

Confirmed cases and tests were divided by ZIP code population in order to obtain a normalized rate per population; due to their right-skew, the natural log of both rate variables was calculated for final use (this also means that ZIP codes with 0 cases or tests were removed in the final analysis, to avoid problems with estimating models with missing values). Percent white, black, and Hispanic are by ZIP code used to classify each ZIP code as either majority-white, majority-Hispanic, or majority-black, based on a level of >50% population of the given racial/ethnic group in that ZIP code. Given the strong historical pattern of segregation in Chicago, these are the key neighborhood indicators of interest, and themselves capture much correlated information on educational attainment, economic opportunity, environmental injustice, healthcare access, etc.

Spatial access to testing locations was calculated using the “access” package in python using a simple “gravity”-based method. This approach calculates the weighted Euclidean distance from each ZCTA centroid to each testing location based on a distance decay parameter of (in this case) -.5 and then sums the total weighted distances from each origin to all testing locations within a given maximum threshold of (in this case) 48,000 meters (or roughly 30 miles). The resulting summed weights were then scaled using a max-min standardization so that the final testing access scores – representing spatial access from a given ZIP code to testing sites within driving distance – ranged from 0 – 100.

3. Methods and Results

This paper takes an exploratory approach to mapping, analyzing, and modeling these data. Maps showing the general spatial patterns of the variables of interest are used to better understand spatial relationships. A correlation analysis also shows the quantitative association between variables of interest without taking into account controlling factors. Finally, two sets of spatial econometric models are estimated using the approach and diagnostics laid out in Anselin and Rey (2014) – the first examine the relationship between neighborhood type and COVID-19 testing rates while controlling for case rates to better understand underlying disparities in testing for COVID-19. The second examine covariates of case

rates while controlling for spatial access to testing sites in order to better understand disparities in the neighborhood burden of COVID-19 in the region.

Figure 1 shows the general spatial patterns for 6 variables of interest; Pearson correlations between all variables of interest in the dataset are shown in Table 2. Several relationships of interest stand out: first, case rate and testing rate are highly-related (.88 correlation), but not perfectly correlated. Both are positively related to the black majority neighborhood dummy and negatively correlated to white neighborhoods, but, interestingly, the case rate is more highly-correlated with Hispanic neighborhoods than testing rate, suggesting that there may be differential testing rates for Hispanic neighborhoods. There is also some overlap between young neighborhoods and those with higher rates of workers in service occupations with Hispanic neighborhoods, while black neighborhoods show higher correlations with service occupations and the proportion of commuters who are transit-dependent. White neighborhoods, on the other hand, are relatively strongly negatively correlated with every variable of interest other than the "high income" urban professional/transit commuting principle component (PC_TPT). The measure of spatial access to testing sites also displays relatively high correlations with case rates and testing rates, but no large relationship with any of the neighborhood types.

TABLE 2. Correlation Table for Variables of Interest

	Log(Case Rate)	Log(Test Rate)	Black	White	Hispanic	PC_OLD	PC_ YOUNG	PC_SRV	PC_TPT	PC_ TRANS
Log(Test Rate)	**0.88**									
Black Neighborhood	**0.50**	**0.47**								
White Neighborhood	**-0.57**	**-0.40**	**-0.52**							
Hispanic Neighborhood	**0.15**	**0.01**	**-0.11**	**-0.41**						
PC_OLD	**0.32**	**0.34**	**0.12**	**-0.13**	**0.11**					
PC_YOUNG	**0.11**	**-0.01**	**0.02**	**-0.34**	**0.34**	**0.17**				
PC_SRV	**0.44**	**0.26**	**0.44**	**-0.64**	**0.34**	**0.12**	**0.29**			
PC_TPT	**0.24**	**0.31**	**0.14**	**0.00**	**-0.07**	**0.32**	**0.34**	**-0.28**		
PC_TRANS	**0.51**	**0.49**	**0.38**	**-0.44**	**0.19**	**0.21**	**0.29**	**0.43**	**0.28**	
Spatial Access to Test Sites	**0.40**	**0.42**	**0.05**	**-0.15**	**0.15**	**0.18**	**-0.01**	**0.09**	**0.14**	**0.33**

In order to better understand whether these relationships are statistically-significant in the presence of control variables – and controls for spatial dependence – two sets of spatial lag models are estimated using the Spatial Two-Stage Least Squares (S2SLS) approach laid out in Anselin and Rey (2014) in GeoDaSpace v.1.2 using White Standard Errors. Spatial lags are appropriate for both sets both from a theoretical standpoint – the vírus is directly transmitted from person to person, and thus we would expect spatial spillovers to occur directly between neighborhoods via this mechanism – and based on standard diagostic tests, namely significant robust Lagrange Multipliers (lag) when applied to Ordinary Least Squares (OLS) estimates of both models with associated spatial weights matrices. The general framework of the S2SLS approach is to treat the spatial lag parameter (**Wy)** from the standard spatial lag model as an endogenous variable that is instrumented for by the first-order spatially-lagged explanatory variables. The standard spatial lag specification is shown in Equation (1).

$$y = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\beta + u, \qquad (1)$$

where y is the dependent variable of interest, $\mathbf{W}$ is a spatial weights matrix (a first-order queen contiguity matrix in this case), $\mathbf{X}$ is a vector of explanatory variables, and u is the error term. To correct for the endogeneity of the spatially-lagged variable, a matrix of instrumental variables $\mathbf{Q}$ is specified in Equation (2).

$$\mathbf{Q} = [\mathbf{X}, \mathbf{WX}]. \quad (2)$$

The first set of models addresses the question of whether or not particular types of neighborhoods demonstrate lower-than-expected testing rates, which is important to understand, given the early stage of the pandemic and the current lack of testing availability. If tests were being conducted without disparity, we would expect the high correlation between testing rate and case rate to apply equally across all three neighborhood types. To examine whether this is the case, three spatial lag models were run with Log(Test Rate) as the dependent variable, Log(Case Rate) as an independent variable – to importantly control for the fact that observations of large or small case rates may in fact be due to the testing occurring within each neighborhood – and each individual neighborhood type.

As the results in Table 3 show, both white and black neighborhoods have a significant positive relationship with testing rate, even after controlling for case rates. However, the results show that Hispanic neighborhoods have a significant negative relationship with testing rate even after controlling for the fact that these neighborhoods may have lower case rates than others. This suggests that there may be some systematic disparity in access to testing for Hispanic neighborhoods, or perhaps a lack of propensity to be tested for residents with uncertain legal immigration status. The spatial lag of the dependent variable is also significant in each model, suggesting that observations of testing rates are spatially-dependent.

TABLE 3. Spatial Two Stage Least Squares Model Results for COVID-19 Testing Rate

Dependent Variable = Log(Test Rate)	White Neighborhoods		Black Neighborhoods		Hispanic Neighborhoods	
Variables	***Coef.***	***Std. Error***	***Coef.***	***Std. Error***	***Coef.***	***Std. Error***
Constant	0.003	0.149	-0.262	0.164	-0.049	0.140
Log(Case Rate)	0.508***	0.060	0.448***	0.054	0.474***	0.051
Neighborhood Dummy	0.106***	0.033	0.065*	0.033	-0.230***	0.041
W_Log(Test Rate)	.337***	0.092	0.344***	0.091	0.352***	0.089
Number of Observations	285		285		285	
Degrees of Freedom	281		281		281	
Pseudo R-squared	0.838		0.833		0.847	

Notes: * denotes estimates significant at p ≤ .1, ** denotes p ≤ .05, and *** denotes p ≤ .01. Instruments for W_Log(Test Rate) are W_ Log(Case Rate) and W_Neighborhood Dummy.

What are the correlates of COVID-19 case rates in the Chicago region? Table 4 shows the results of the second set of models, which estimate the relationship between the transportation, occupation, and racial/ethnic covariates of interest with observed case rates. The progression from OLS model results to spatial lag results uncovers some interesting information about the relationship between the variables examined here. In the initial OLS model, all variables other than the Hispanic neighborhood dummy are significant, suggesting that black neighborhoods and those with larger proportions of older residents, workers in service occupations, and commuters using transit (whether transit-dependent or not) are all more likely to demonstrate higher case rates, even when controlling for spatial access to testing sites.

However, when the spatial dependence parameter is added into the second model, both transit variables (as well as the principle component denoting higher proportions of younger residents) become insignificant, suggesting that at least some of the positive association between transit ridership and increased COVID-19 case rates is in fact due to the spatial spillover of case rates, i.e., the proximity of these neighborhoods to other neighborhoods with high case rates explains more variation in observed case rates than the propensity of transit use in these neighborhoods. Three features remain significant predictors of COVID-19 case rates, though, even after controlling for spatial dependence and spatial access to testing sites: black majority neighborhoods, neighborhoods with higher proportions of older residents, and neighborhoods with higher proportions of service workers in essential occupations.

TABLE 4. OLS and Spatial Two Stage Least Squares Model Results for COVID-19 Case Rate.

Dependent Variable = Log(Case Rate)	OLS Estimation		Spatial Lag Estimation	
Variables	***Coef.***	***Std. Error***	***Coef.***	***Std. Error***
Constant	-6.488***	0.098	-2.690***	0.590
Black Neighborhood Dummy	0.551***	0.108	0.366***	0.081
Hispanic Neighborhood Dummy	0.171	0.113	0.051	0.090
PC_OLD	0.133***	0.045	0.077*	0.041
PC_SRV	0.138***	0.033	0.066**	0.033
PC_TPT	0.121***	0.035	0.011	0.039
PC_TRANS	0.122**	0.056	0.053	0.073
PC_YOUNG	-0.078***	0.029	-0.015	0.028
Spatial Access to Testing Sites	0.008***	0.002	0.004***	0.001
W_Log(Case Rate)			0.597***	0.093
Number of Observations	284		284	
Degrees of Freedom	275		275	
Adjusted/Pseudo R-squared	0.512		0.709	

Notes: * denotes estimates significant at p ≤ .1, ** denotes p ≤ .05, and *** denotes p ≤ .01. Robust LM value for OLS specification = 14.1 w/ .0002 probability. Instruments for W_Log(Case Rate) are W_Black Neighborhood Dummy, W_Hispanic Neighborhood Dummy, W_PC_OLD, W_PC_SRV, W_PC_TPT, W_PC_TRANS, W_PC_YOUNG, and W_Spatial Access to Testing Sites.

4. Discussion and Conclusion

The results of this analysis, while exploratory, provide some useful information about the current socio-demographic disparity in COVID-19 testing and (observed) infection rates in the Chicago, IL region. First, there appears to be a disparity in COVID-19 testing accessibility in Hispanic majority neighborhoods, even when controlling for the observed case rates in these neighborhoods. Whether this represents a systematic disparity in physical (transportation-related) access to testing locations, a propensity to avoid testing for those with uncertain legal immigration status, or other social, cultural, or economic reasons, remains to be seen. However, this observed disparity could have consequences in the future course of the pandemic if Hispanic neighborhoods become more heavily-affected by the virus at a later time. These findings could also suggest that actual rates of infection in Hispanic neighborhoods are higher than currently reported, creating an important information gap in our understanding of the virus' spread and ultimate toll.

Second, three primary features appear to significantly predict observed COVID-19 case rates, even when controlling for spatial spillovers in infection activity and spatial access to testing sites: black majority neighborhoods, neighborhoods with larger proportions of older (65+) residents, and neighborhoods with higher proportions of workers in service occupations (including healthcare service, protective service, food preparation, cleaning service, and transportation workers who perform essential jobs that are not able to stay at home during the pandemic) all have a significant positive relationship with COVID-19 infections. From this analysis, it does not appear that high proportions of transit use are significantly related to infection rates beyond the spatial spillover characteristics of infections. While these findings match the existing hypotheses about COVID-19 risk, they also underscore the important spatial and racial/ethnic disparity of COVID-19's burden, as these neighborhoods also tend to be the most lacking in economic and political capital and access to quality healthcare.

While these data are certainly preliminary – as we have no knowledge about the ultimate course of the pandemic – if these findings continue to hold, they suggest that 1) in the short term, additional resources need to be deployed to minority-majority neighborhoods, both to increase access to testing and to enhance social distancing and other protective public health measures, and 2) in the long term, policies to effect significant structural change in these neighborhoods should be pursued to help majority-minority

neighborhoods (and thus our society as a whole) become more resilient to future pandemics and other public health crises.

5. Figures

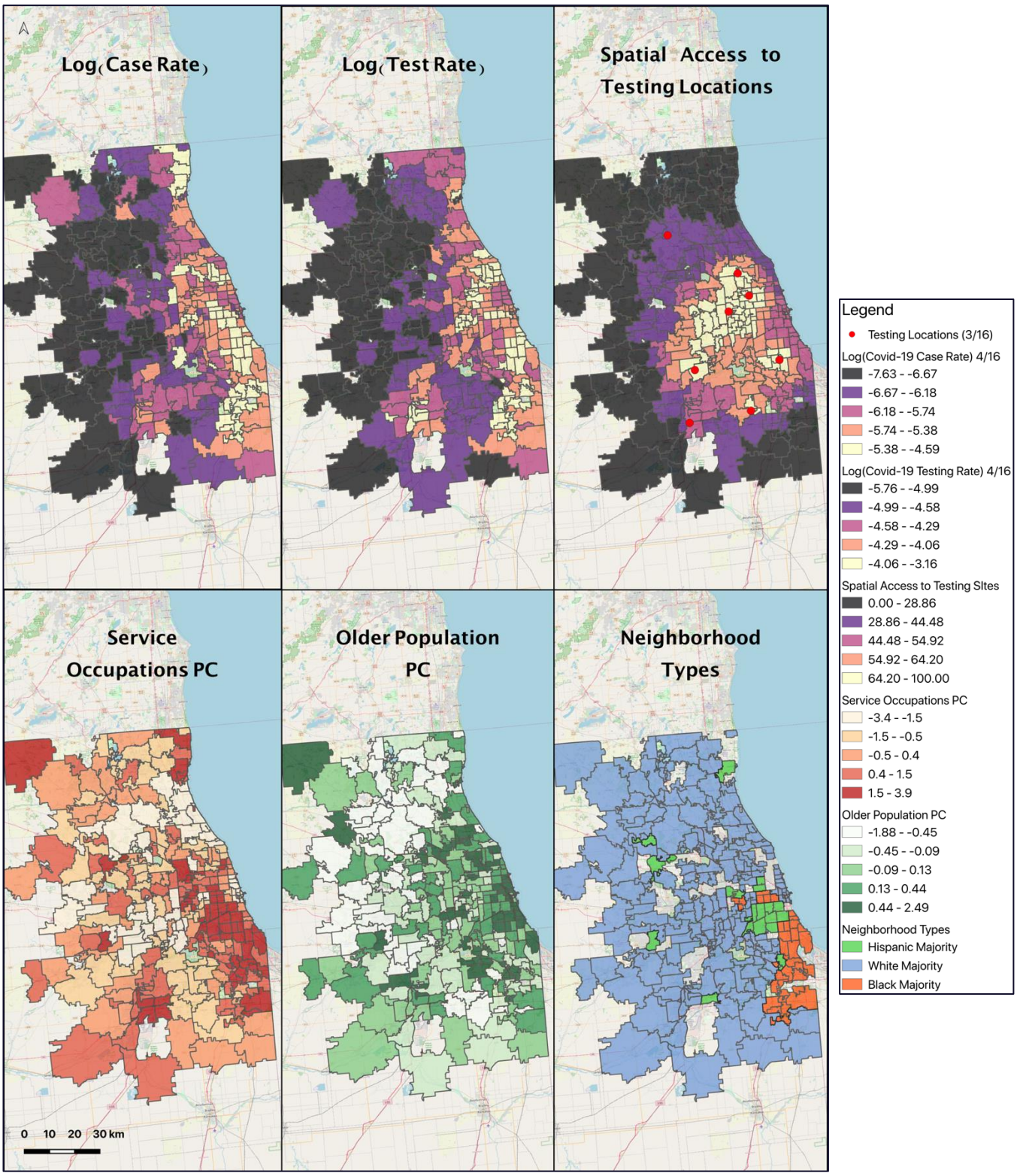

Figure 1. Spatial Patterns for Variables of Interest in Chicago, IL Region

REFERENCES

Anselin, L., & Rey, S. (2014). Modern Spatial Econometrics in Practice: A guide to GeoDa, GeoDaSpace, and PySAL. Chicago, IL: GeoDa Press.

Arias, E., Escobedo, L. A., Kennedy, J., Fu, C., & Cisewski, J. (2018). U.S. Small-area Life Expectancy Estimates Project: Methodology and Results Summary. National Center for Health Statistics. *Vital Health Statistic, 2*(181). URL: https://www.cdc.gov/nchs/nvss/usaleep/usaleep.html.

Braveman, P. A., Cubbin, C., Egerter, S., Williams, D. R., & Pamuk, E. (2010). Socioeconomic disparities in health in the United States: What the patterns tell us. *American Journal of Public Health.* DOI: 10.2105/AJPH.2009.166082.

Center for Spatial Data Science (CSDS) (2020). U.S. COVID-19 Atlas. University of Chicago. URL: https://geodacenter.github.io/COVID/.

Chicago Department of Health (CDPH) (2020). Latest Data. URL: https://www.chicago.gov/city/en/sites/COVID-19/home/latest-data.html.

CIA Factbook (2017). The World Factbook - Country Comparison: Life Expectancy at Birth. URL: https://www.cia.gov/library/publications/the-world-factbook/rankorder/2102rank.html.

COVID-19 Drive Thru Location Tracking (2020). COVID-19 Drive Thru Location Tracking. URL: https://docs.google.com/spreadsheets/d/1svnaZ2UG_ryFr8jjqVx7ZVZksBue4EQUJ4dolMDJx70/edit#gid=0.

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases.* https://doi.org/10.1016/S1473-3099(20)30120-1.

Illinois Department of Health (IDPH) (2020). Coronavirys Disease 2019 (COVID-19): COVID-19 Statistics. URL: http://www.dph.illinois.gov/COVID19/COVID19-statistics.

Johns Hopkins University (JHU) (2020). COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. URL: https://coronavirus.jhu.edu/map.html.

Kolak, M., Bhatt, J., Park, Y. H., Padrón, N. A., & Molefe, A. (2020). Quantification of Neighborhood-Level Social Determinants of Health in the Continental United States. *JAMA Network Open, 3*(1). DOI: 10.1001/jamanetworkopen.2019.19928.

National Center for Health Statistics (NCHS) (2015). US Small-area Life Expectancy Estimates Project (USALEEP). National Vital Statistics System. URL: https://www.cdc.gov/nchs/nvss/usaleep/usaleep.html.

Solar, O., & Irwin, A. (2010). A conceptual framework for action on the social determinants of health. *Social Determinants of Health Discussion Paper 2 (Policy and Practice)*. Geneva, Switzerland: World Health Organization (WHO). URL: https://www.who.int/sdhconference/resources/conceptualframeworkforactiononsdh_eng.pdf.

Williams, D. R., Mohammed, S. A., Leavell, J., & Collins, C. (2010). Race, socioeconomic status, and health: complexities, ongoing challenges, and research opportunities. *Annals of the New York Academy of Sciences,* DOI: 10.1111/j.1749-6632.2009.05339.x.

Williams, D. R., Priest, N., & Anderson, N. (2016). Understanding Associations between Race, Socioeconomic Status and Health: Patterns and Prospects. *Helath Pyshcology* 35(4): 407-411.

Worldometer (2020). COVID-19 CORONAVIRUS PANDEMIC. URL: https://www.worldometers.info/coronavirus/.