COVID INFECTION INDEX: A GEOGRAPHICALLY WEIGHTED APPROACH IN ASSESSING COVID INFECTION IN NYC

By Alex Deleon

DECEMBER 20, 2020
HUTNER COLLEGE

Abstract:

The aim of this study is to identify which variables exhibit the strongest correlation with the infection rate of COVID-19, specifically in relation to occupation and transportation, household and living conditions, socioeconomic status, everyday activities, and health and age. Additionally, the Covid Infection Index (CII) is developed based on the identified variables, which serves as a tool to assess the susceptibility of communities to COVID-19 infection, while accounting for the diverse geographical landscape of New York City (NYC). Multiple linear regression (MLR) and geographical weighted regression (GWR) were used to measure the relationship between the variables considered and the infection rate of COVID-19. The GWR analysis was run using Zip Code Tabulation Areas. Geographically weighted principal component analysis (GWPCA) and principal component analysis (PCA) are then used to create the CII using Census Tract geography and data. Given the granularity of Census Tracts and the use of geographic weighted techniques, this study provides a method of understanding how the impact of COVID-19 individually affects the diverse communities of NYC.

Introduction

COVID-19 (COVID) is an infectious disease that is transmitted through respiratory droplets or aerosols[1]. Any setting where people interact with each other poses a risk of spreading the infection. Those who regularly interact with other people and who are ill-prepared or disadvantaged are more at risk. The question then becomes, what provokes people to interact and what groups of people are ill-prepared or disadvantaged? We will approach this problem from many angles, hoping to capture some essence of the answer to this difficult question.

Given the nature of the disease, interaction will be treated as any size (small or large) group of people meeting in an indoor or outdoor setting who speak or engage with one another. Here we will discuss interaction due to occupation and transportation, household and living conditions, socioeconomic status, everyday activities, and health and age.

Occupation and Transportation

Essential workers are at the forefront of the pandemic. Preventing the spread among them is largely dependent on mitigation measures. Employers are required by law to provide personal protective equipment and information on COVID cases or potential exposures at the workplace. Despite mitigation measures and employer transparency, essential workers are not removed from the risk. They report to work at common settings: childcare workers report to schools or institutions with students and staff; retail workers interact with other employees and patrons per business demands; manufacturers report onsite at factories or workstations—all of which involve interacting with others in some way.[2] [3]

How these employees get to work may also be a concern. It was noted in a New York Times article (NYT)[4] that Metropolitan Transportation Authority (MTA) ridership decreased by nearly 90% at the start of the pandemic. And that even with NYC's containment of the virus, ridership is only 20% of what it used to be. With these numbers, it is difficult to say if ridership—as it is now—is contributing to COVID-19 transmission.

Sam Schwartz, in an extensive report on public transit and COVID, among other things, defends the point that there is no correlation between public transit and COVID infection.[5] Correlation in this report is not to be confused with the Pearson correlation coefficient. It should rather be taken as meaning a notable relationship or connection between the two. Schwartz defends this point by graphing 7-day rolling averages of infection counts and transit ridership of various cities, often illustrating an inverse relationship or no relationship at all. He additionally provides alternative explanations for COVID infection in these various cities (as linked to family gatherings, prisons, meat-packing plants, careless youth, hasty re-openings, etc.). Also, he provides comprehensive lists and descriptions of strategies each city's transit system is using to combat the virus, and the positive effects of these strategies.

Ridership, as he also notes, has plummeted globally. It is suggestive to say that public transportation has no correlation to COVID infection if ridership has practically been inactive and taken utmost precautions against the virus. It is true that there are many causes for the spread of COVID. But this gives no reason to suggest that public transit with all precautions imaginable, operating at hypothetically 75% capacity pre-pandemic, will have no correlation to COVID infection—as Ben Yakas, a writer from Gothamist, in citing Schwartz has alluded to.[6]

Another form of transportation for employees that may be a concern is carpooling. Many have provided guidance on carpooling, acknowledging the potential risk.[7] [8] New York University's School of Global Public Health has developed COVID trainings for rideshare drivers, aiming to instruct drivers of best practices in mitigating exposure to COVID. This is an important initiative for drivers given how difficult it is that maintain distance from passengers.[9] Essential workers and drivers alike are faced with the risk of contracting COVID at the workplace or in transit. Experiencing the same risk are those residing in crowded homes or group quarters.

Household and Living Conditions

In a cross-sectional study done on 400 pregnant women, overcrowded homes were found to be a major predictor of COVID spread.[10] [11] Overcrowding is a stark reality for many New Yorkers. According to a report by Scott M. Stringer, in 2013 "there were over 272,000 crowded dwelling units in New York City, occupied by more than 1.46 million residents.[12] These numbers in conjunction with our earlier example of a household risk raises enormous concern. But residents are not the only ones affected by shared space. People in group quarters face a similar risk.

The census definition of group quarters varies significantly throughout history. Before 1930, institutions, hotels, and boarding houses were categorized as large houses.[13] Today group quarters are defined to mean a range of things, for example: federal and state prisons, local jails, student housing, psychiatric hospitals, group homes, nursing facilities, and so on.[14]

Local prisons and jails have been a hotpot for COVID spread. At the time of study, it was found that the rate of COVID in local jails and prisons in Massachusetts "was nearly 3 times that of the Massachusetts general population and 5 times the US rate."[15] Cases are also widespread throughout universities, infecting thousands of students across the US.[16] In nursing homes, (essential) workforce and community spread led to alarming infection and death rates.[17] These are only three examples of the risk in group quarters.

Socioeconomic Status

In this study, socioeconomic status is defined according to the American Psychological Association (APA) as a composition of three dimensions: education, income, and occupation.[18] Higher education improves critical thinking, problem solving and decision making, and

navigating daily life.[19] And as it is commonly known, higher education leads to greater income, satisfaction with occupation, and overall an easier life.[20] We can see that the dimensions of socioeconomic status are interrelated, often one is not without the other.

Low socioeconomic status prevents access to basic resources such as proper health care, nutritional food, and technology.[21] [22] Such an important issue is only exasperated in a raging pandemic: Employment has dried up, food insecurity has risen, and mental health has declined.[23] [24] Everyday activities such as food shopping, dining, religious congregation, doing laundry, getting a haircut, or meeting for coffee now all come with a risk. Public spaces where people commit to everyday activities are often called Points of Interest (POI). [25] [26]

Points of Interest

POI like churches, synagogues, or mosques have been linked to COVID outbreaks. [27] [28]Abrahamic religions like Christianity, Judaism, and Islam require once a week some form of devotion and observance to their beliefs. For Christianity, that is Sabbat (or Sabbath) on Saturday or Sunday; and for Judaism, Sabbat is from Friday to Saturday. [29] [30]During Sabbat, people pray, recite, socialize, and interact together. On Fridays (Jummah), it is required in Islam that all men congregate, listen to sermon (khutbah), and pray together. [31] Considering this, it is clear why COVID is rampant in places of worship. People of faith are trying to fulfill their religious obligations while weighing the risk of COVID.

For others, it may be POI like barbershops that provoke the spread of COVID. Barbershops and salons in black and Latino communities are often treated as trusted social settings. [32] Aware of this, the Colorado Black Health Collaborative (CBHC) Barbershop/Salon Health Outreach Program has even utilized barbershops as a scene for preventative health care and pedagogy.[33] But what makes this program so effective—the trust-based and social atmosphere—may facilitate COVID spread.

Other POI that this study will consider are liquor stores, retail food stores, and laundry centers. Although liquor stores are not necessarily a place of congregation, the number of liquor stores in an area makes alcohol more accessible, which in turn may promote social drinking and overall alcohol consumption. [34] Greater alcohol accessibility is also associated with unemployment, poor financial well-being, and underage drinking, tapping into other related issues of COVID.

Today grocery shopping can be done online. Many have taken up this opportunity as a way of circumventing the virus, while others do not have this luxury. [35] New Yorkers who receive Supplemental Nutrition Assistance Program (SNAP) benefits are limited in their selection of online grocery shopping[36], with only three stores accepting online SNAP usage on selected items—not to mention that SNAP does not cover any delivery charges. This means that SNAP

beneficiaries are more likely to visit local markets or food stores, expanding their selection and avoiding delivery charges.

As for laundry centers, visiting them is generally safe so long as safety precautions are taken by both patrons and business owners, and face-to-face interactions are limited. [37] Whether these conditions are followed is largely dependent on the business owners, the area, the number of patrons, and personal choice.

Health and Age

The senior population and those with pre-existing health conditions are more likely to have compromised or weaker immune systems resulting in a higher severity of symptomology and potential fatality. [38] Given this, it is expected these group would take extra precautionary measures in handling the pandemic. However, conditions like high blood pressure, high cholesterol, and smoking are prevalent among certain essential worker and low socioeconomic groups.[39] [40] And as discussed, these are a more susceptible to COVID.

Many seniors are exercising extreme caution, and as a result, are facing isolation and depression. [41] [42] Low-income seniors face additional challenges such as food insecurity and crowded housing. [43] [44]

About the Data

COVID as we have outlined is a multifaceted problem, much of which we have not considered. Many have investigated other important issues in relation to COVID such as racial disparity and environmental factors. [45] [46]

However, this study will only investigate the variables discussed thus far. Outlined in table 1 below are the variables this paper will consider, data sources, and a brief description of the processing involved in obtaining each dataset (if any). In gray are the variables that were removed after variable selection.

	Acronym	Description	Source	Processing
1	**OVER 65**	**% over 65**	**ACS_DT**	**DP02_0076PE**
2	**DISABL**	**% with a disability**	**ACS_ST**	$\frac{\mathbf{B18106\_001E}}{\boldsymbol{S1810\_C01\_001E}}$
3	**NOINSUR**	**% without insurance**	**ACS_DT**	**DP03_0099PE**
4	**ESS_WRKER**	**% essential workers[1]**	**ACS_ST**	$\frac{\boldsymbol{Tot}\ \mathbf{PTM} + \boldsymbol{NCM} + \boldsymbol{S} + \boldsymbol{H}}{\boldsymbol{Tot\ Employed}}$,
5	**AVG HH**	**Average person per household**	**ACS_ST**	
6	**RENT 35%**	**% whose rent makes up 35% of income**	**ACS_DT**	**DP04_0142PE**
7	**GROUP Q**	**% in group quarters**	**ACS_ST**	$\frac{\mathbf{B26001\_001E}}{\boldsymbol{S1810\_C01\_001E}}$
8	**HIGHCH**	**Crude rate of Pop with High cholesterol**	**Health.data.ny.gov**	
9	NOCOMP	% without access to a computer	ACS_ST	$\frac{\mathrm{S2801\_C01\_011E}}{S2801\_C01\_001E}$
10	NOINTER	% without access to internet	ACS_ST	$\frac{Tot\ HH\ without\ Internet}{Tot\ HH} = \frac{\mathrm{S2801\_C01\_019E}}{S2801\_C01\_001E}$
11	NOHSDP	% 25 and over without hs diploma	ACS_DT	DP02_0060PE
12	BELPOV	% below poverty	ACS_DT	DP03_0119PE
13	PUBTRANS	% employed who take public transportation	ACS_ST	$\frac{Tot\ Public\ Trans\ Emp}{Tot\ Employed} = \frac{\mathrm{S0801\_C01\_009E}}{S2401\_C01\_001E}$
14	CARPOOL	% who carpool to work with 4+ people	ACS_ST	$\frac{Tot\ Carpool\ Emp}{Tot\ Employed} = \frac{\mathrm{S0801\_C01\_007E}}{S2401\_C01\_001E}$
15	NOVWH	% who do not own a vehicle	ACS_DT	DP04_0058PE
16	LIQSTORE	# of liquor stores[1,8]	Licensing Bureau and Information Technology - GIS	Query for Liquor stores in NYC
17	BARBSAL	# of barbershops and Salons	NYC OPEN DATA	Direct Download
18	FOODST	# of retail food stores[10,12]	Dept. of Agriculture and Markets	Direct Download
19	LAUND	# of laundry centers	NYC OPEN DATA	Direct Download
20	WRSHP	# places of worship	MAP PLUTO Buildings Dataset	query for building class M, places of worship or religious congregation
21	ASLIVNH	Number of beds in nursing homes and assisted living homes	health.data.ny.gov	
22	HIGHCH	Crude rate of Pop with High cholesterol	Health.data.ny.gov	
23	SMOKE	Crude rate of chronic smokers	Health.data.ny.gov	
24	BINGE	Crude rate binge drinkers	Health.data.ny.gov	
25	HIGHBP	Crude rate pop with high blood pressure	Health.data.ny.gov	
26	OCCUN	% in occupied housing units with 1.01 or more persons per room	ACS_ST	$\frac{\mathrm{S2501\_C01\_007E} + \mathrm{S2501\_C01\_008E}}{S2501\_C01\_001E}$

Table 1. Variables bolded are the ones that will be used in the analysis after variable selection. The variables grayed out are those that were removed during variable selection.

[1] ESS WRKER are considered those who work in production, transportation, and material moving occupations (PTM), in natural resources, construction, and maintenance occupations (NCM), in protective service occupations (S), or as healthcare practitioners and technical occupations (H). $PTM = \mathrm{S2401\_C01\_033E}$, NCM = S2401_C01_029E, H = S2401_C01_015E, and Total Employed = S2401_C01_001E

As we will see, data must be available both at the Zip code Tabulation Area (ZCTA) and census tract level. These entities differ widely in shape, purpose, and population size. [47] ZCTA are much larger, serving as postal service areas, with population sizes as low as 1,400 and as high as 100,000. [48] Census tracts are small statistical areas created and used by the Bureau of Census (CENSUS) to analyze the population. Census tracts maintain relatively uniform population sizes, ranging from 1,000 to 8,000 people. Census widely disseminates data at both these geographies, and so majority of the data selected is from 2018 American Community Survey Detailed and Subject Tables (ACS DT, ACS ST).

Various POI (list the POI) datasets were collected from NYC or NY Open Data. These datasets serve as a proxy for measuring how often or how much people visit them, since data like this was not free to the public.

Chronic health data was collected from the Center for Disease Control's 500 Cities Project. This project provides estimated chronic health data of 27 health variables at various geography levels such as census tracts, zip codes, and counties.

Scope of Study

- What variables most account for COVID infection rates in NYC?
- Using these variables, how can we create a composite indicator to capture the "overall" vulnerability of communities in NYC?
- In addressing the two questions above, how can we account for the geographic heterogenous impact of COVID on NYC communities?

Approach

To answer the first question, multiple linear regression (MLR) analysis will be used as a means of selecting the variables that contribute most to the measurement of COVID infection rates in NYC; for the second, variables remaining from this selection will then be used to create a composite indicator using principal component analysis (PCA); and lastly, for the third, the same analyses will be performed using geographically weighted regression (GWR) and geographically weighted PCA (GWPCA).

Because the zip code is the smallest geography for which COVID data is available, MLR and GWR will be run at the *zip code* level. This will allow us to investigate the direct relationship between each variable on the spread and impact of COVID. PCA and GWPCA will then be performed at the *census tract* level to provide a more granular understanding of which communities are impacted.

MLR Model. In the MLR model, the infection rate of each NYC *zip code* as our **Y** (dependent) variable will be regressed against several **X** (explanatory) variables. This model does not consider the spatial properties of each zip code, that is, the location or proximity of each zip code from each other. Explanatory variables will be tested for collinearity and response variables will be tested for normality. Stepwise regression will be used to select the variables that account for most of the variance in **Y**. Despite being removed after auto-selection, some

variables will be kept for theoretical reasons. Residual diagnostics such as residual plots and heteroscedasticity tests will be iteratively used to assess the quality of the model.

GWR Model. Once the MLR model is refined, a geographically weighted regression[2,3,11] (GWR) will be used with the explanatory variables selected through the above step. This model is a localized version of MLR, one that considers the spatial properties and heterogeneity of NYC zip codes. Moran's I will be used to test for spatial autocorrelation. Like MLR, localized VIFs will be used to assess multicollinearity.

Grouping Variables. To prepare for PCA and GWPCA, variables will be grouped into subthemes by considering the conceptual relationship between one another. For instance, before analysis, it would be reasonable to group the essential worker variables (5-7) into the same subtheme.

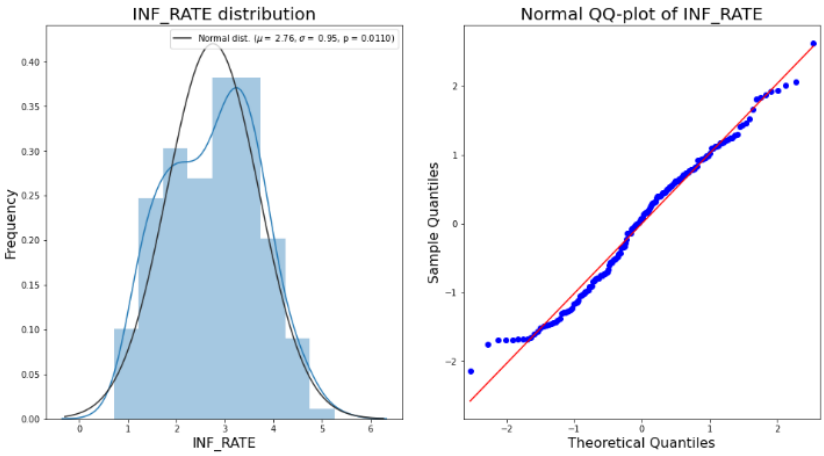
PCA. To perform PCA, we will first transform the original values of each variable to "standard scores" by subtracting the mean and dividing by the standard deviation. Principal component analysis (PCA) is then to be conducted for each subtheme. This means that the location and distance of each census tract from each other will not be considered in this analysis. For each subtheme, we calculate the sub-index as a weighted sum of the standardized scores of its variables, where the weights are the loadings of the corresponding variables under the 1st PCA component. PCA will then be run again on the subindices to create the total index itself. Specifically, we standardize the subindices, then run PCA on them, and then calculate the weighted sum.

GWPCA. After running PCA, we will run its geographically weighted counterpart—geographically weighted PCA (GWPCA)[7,9]. Much like PCA, we will compute a subindex for each subtheme using the first component loadings. Thereafter we will run GWPCA on the subindices and compute the total index (the overall COVID vulnerability score).

Both GWR and GWPCA will be carried out using the GWmodel R package[5].

MLR Model Assumptions

Shapiro Wilks is used to test the response variable COVID-19 infection rates (total cases over total population as INF_RATE) for normality. QQ and frequency plots were used as a visual aid for this test. INF_RATE failed the Wilks test with p-values less than 0.05. Judging for the QQ-



plots, outliers are present. However, the model is robust to the moderate violation of normality given the large sample size (n =187).

Figure 1. Frequency plot of QQ plot of INF_RATE

Multicollinearity was assessed using variance inflation factors (VIFs) and correlation matrices. In some cases, if variables demonstrated high VIFs (typically above 5) they were removed or combined with another to reduce collinearity. Variables that demonstrated unstable and uninterpretable coefficients were also removed. The following variables that were removed are listed in table 2, along with their reasons for removal.

Variable	Reason(s) for removal
NO INTER	• Highly correlated with % of households without a computer • Negative and insignificant coefficient. • Large confidence interval.
NOHSDP	• Highly correlated with % of "essential workers"
BELPOV	• Highly correlated with % of "essential workers"
PUBTRAN	• Highly correlated with % of "essential workers" and % of people who do not own a vehicle • Negative and insignificant coefficient. • Large confidence interval
NOVEH	• Highly correlated with % of "essential workers" and % of people who do not own a vehicle
CARPOOL	• Highly skewed variable concentrated around 0%-10%, resulting in insignificant coefficient.
OCCUN	• Uninterpretable negative coefficients. Overcrowding is expected to lead to higher infection rates. This variable did not capture this. Average number of persons in a household did instead.
POI: WORSHP, LIQ, LAUND, BARBSAL, FOODST	• Variables resulted in uninterpretable negative coefficients. One would expect that the coefficients to be positive since points of interest are places of congregation. But the number of points of interest may not be relevant to how often--and how many--people visit them.
SMOKE	• Highly correlated with high cholesterol
HIGHBP	• Highly correlated with high cholesterol and chronic smoking
BINGE	• Negative and insignificant results • Oddly, this variable was not found to be correlated to many of the other health variables. It may be possible that many binge drinkers do not report this information.

Table 2. Variables that were removed after selection.

MLR Model

According to table 3, of the 8 variables remaining after variable selection, *ESS WRKER* and *OVER 65* are not significant. *GROUP Q* is moderately significance with *p*=.05. *RENT 35%* and *NO INSUR* demonstrate high significance with *p*<.05. while AVG HH and DISABL are extremely significant with *p*<.001. The model accounts for approximately 61% of the variance in INF_RATE.

Variable	*B*	*SE*	*t*	*p*	*[0.025*	*0.975]*
Intercept	*-1.93*	*0.412*	*-4.691*	*0.000****	*-2.743*	*-1.118*
DISABL	*0.0804*	*0.017*	*4.620*	*0.000****	*0.046*	*0.115*
OVER 65	*-0.023*	*0.018*	*-1.290*	*0.199*	*-0.058*	*0.012*
NO INSUR	*-0.037*	*0.018*	*-2.126*	*0.035***	*-0.072*	*-0.003*
RENT 35%	*0.0234*	*0.008*	*2.782*	*0.006***	*0.007*	*0.040*
GROUP Q	*0.0286*	*0.014*	*1.971*	*0.050**	*0.000*	*0.057*
AVG HH	*0.6649*	*0.151*	*4.393*	*0.000****	*0.366*	*0.964*
ESS WRKER	*0.0078*	*0.005*	*1.468*	*0.144*	*-0.003*	*0.018*
HIGH CHOL	*0.0479*	*0.027*	*1.766*	*0.079**	*-0.006*	*0.101*
R^2	*0.627*	*F*	*35.77*			
Adj R^2	*0.612*	*P (F)*	*1.42E-32*			
n	*177*					
****p<.001*	***p<.05*	**p<.1*				

Table 3. MLR results.

OVER 65 and NO INSUR are negatively related to INF_RATE. The well-known risk associated with old age may explain why the coefficient is negative. The senior population may be taking extreme precautions to avoid infection. The OVER 65 coefficient is interpreted as for every 25% increase in population over 65, there is a 0.5% decrease in the COVID infection rate. Given the range INF_RATE (from 1% to 5%), a 0.5% decrease is relatively high. Those with no insurance may feel less inclined or unaware of their access to testing. In other cases, those without insurance may not be a citizen or foreign born, lacking basic documents required for free testing like a social security or identification card. From table 3, for 25% increase in population with no insurance, there is a 0.75% decrease in COVID infection.

AVG HH and DISABL both have high positive relationships to INF_RATE. The effect overcrowding seems to be captured by AVG HH, suggesting that on average for every 1 person per household, there is approximately a .70% increase in INF_RATE. AVG HH values in NYC are as high as four per household for some ZCTA, implying a 2.8% increase in INF_RATE.

DISABL is related to 6 difficulty types: self-care, independent living, ambulatory, cognitive, hearing, and vision difficulty. [49] If a person possesses one or more of the six difficulty types, they are considered as having a disability by CENSUS. People within this group have trouble

with Basic or Instrumental Activities of Daily Living, likely requiring regular help from family, friends, nurses, or home care workers. [50] [51]

This level of daily interaction may explain why the coefficient is so positively high, suggesting that for every 12% increase in DISABL, there is 1% increase in INF_RATE.

GWR Model

Prior to running GWR, spatial auto correlation and local multicollinearity were tested for. Using 999 simulations, a (Monte Carlo) permutation test on Moran's I was used to assess spatial autocorrelation. An adaptive kernel was used to reduce unstable or locally multicollinear results, ensuring that each local model has the same number of samples, as opposed to a distanced based kernel which may lead to low-sample local models.

From the summary table 4, the variables vary spatially. 5 of the MLR variables from table 4 can be found near the median values of their respective local coefficients. NO INSUR is an exception, falling under 1st quantile. The global adjusted R-squared improved by nearly 11%, from 61% to 72%.

Variable	*Min*	*1st Qu*	*Median*	*3rd Qu*	*Max*	*B*	*Between*
Intercept	*-5.014*	*-2.137*	*-1.729*	*-0.767*	*3.169*	***-1.9304***	*1st to Med*
DISABL	*0.001*	*0.029*	*0.049*	*0.092*	*0.128*	***0.0804***	*Med to 3rd*
OVER 65	*-0.062*	*-0.029*	*-0.003*	*0.024*	*0.065*	***-0.0227***	*1st to Med*
NO INSUR	*-0.081*	*-0.029*	*-0.013*	*0.004*	*0.027*	***-0.0372***	*Min to 1st*
RENT 35%	*0.002*	*0.018*	*0.034*	*0.045*	*0.053*	***0.0234***	*1st to Med*
GROUP Q	*-0.021*	*-0.004*	*0.022*	*0.055*	*0.125*	***0.0286***	*Med to 3rd*
AVG HH	*0.094*	*0.308*	*0.496*	*0.670*	*0.925*	***0.6649***	*Med to 3rd*
ESS WRKER	*-0.011*	*0.002*	*0.005*	*0.010*	*0.025*	***0.0078***	*Med to 3rd*
HIGH CHOL	*-0.142*	*0.001*	*0.023*	*0.064*	*0.188*	***0.0479***	*Med to 3rd*
R^2	*0.79*						
Adj R^2	*0.72*						
n	*177*						

Table 4. GWR summary table.

Figure 2 shows that the local R-squared is much higher in Manhattan and Staten Island, ranging from .85 to .95. In west Brooklyn, some neighborhoods close to these boroughs like Fort Hamilton Sunset Park, and park slope have similar R-squared values. The Bronx and east Brooklyn—near neighborhoods like Sheepshead Bay and Canarsie—R-squared values value with .75 and .85. R-squared values are at their lowest in queens, from .65 to .75.

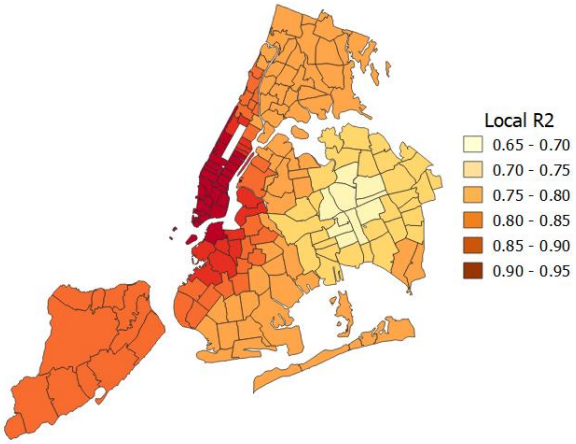


Figure 2. Local R-squared values from GWR

The local models seem to display a general pattern, some of which will be discussed using maps of the local coefficients. For simplicity, two shades of colors were used to represent negative and positive values. Shades of orange represent positive values and shades of green represent negative values. To make the comparison simpler, Jenks breaks with 5 classes were used for each local coefficient map. Alongside each coefficient map, a map of the local t-values was created to assess local significance. T-values with absolute values greater than 1.96 or 2.58 can be viewed as 95% and 99% significance levels.

Figure 3 shows that most of the high local coefficients of DISABL are concentrated in East Queens, where there is about a 1.1% to 1.3% increase in INF_RATE for each 10% increase in DISABL. This coefficient values lessens as zip codes near Manhattan, where less than .30%, or .30% to .50%, increase is seen. Similar patterns can be seen in AVG HH, GROUP Q, ESS WRKER, and OVER 65 local coefficient maps in Appendix A.

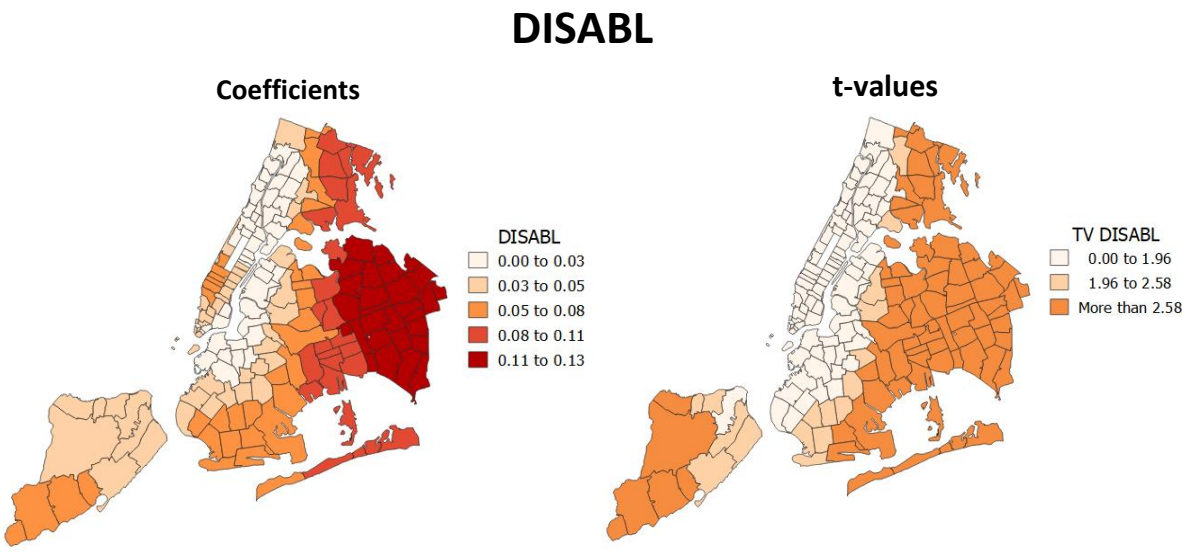


Figure 3. Local DISABL coefficients to left and corresponding local t-values to the right

NO INSUR demonstrates a relationship, unlike the other variables. In Figure 4, the majority of the ZCTA possess negative coefficients. For the ZCTAs do possess a positive coefficient, they are near 0 and are insignificant. This abnormality should not be taken as suggesting that NO INSUR results in lower INF_RATE(s). Theoretically, COVID transmission remains the same for NO INSUR and those with insurance. This may suggest that people with insurance are, for different reasons, not getting tested; and that this abnormal pattern persists throughout NYC.

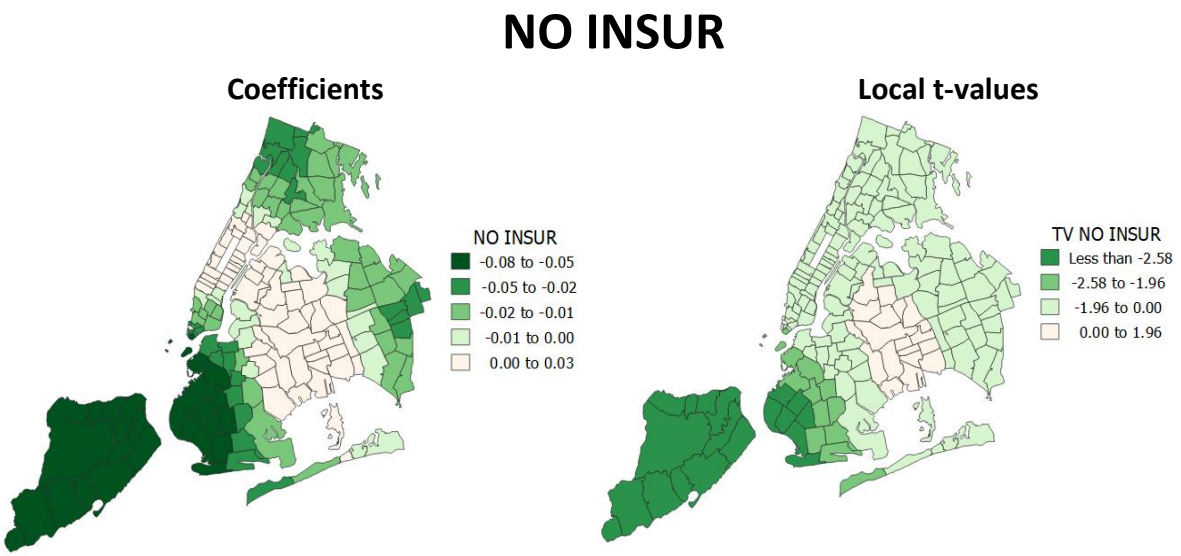


Figure 4. Local NO INSUR coefficients to left and corresponding local t-values to the right

RENT 35% demonstrates the inverse of the primary pattern: high positive coefficients are concentrated in ZCTAs in or near Manhattan. In Manhattan and neighboring ZCTA, for every 20% increase in RENT 35%, there is a .60% to 1.00% increase in INF_RATE.

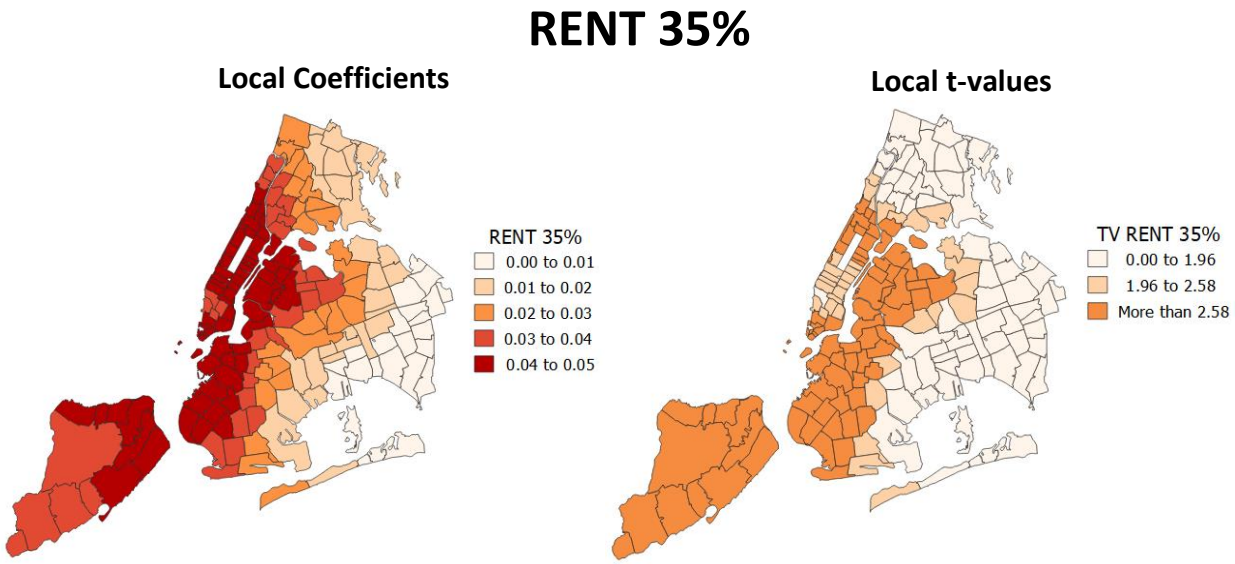


Figure 5. Local RENT 35% coefficients to left and corresponding local t-values to the right

PCA Analysis

The remaining 8 variables were grouped into sub-indices that most captured their conceptual relationships. Since majority of the transportation-related were removed, ESS WRKER was grouped in the Socioeconomic subindex (SOCIOECON)

Formally, ESS WRKER, RENT 35%, and NO INSUR were grouped into SOCIOECON. OVER 6, HIGH CHOL, and DISABL were grouped into Health and Age subindex (HLTHAGE). AVGHH and GROUP Q were grouped into Household and Living Conditions subindex (HSELIV). In discussing the PCA analysis, principal components will be referred to as PC. Additionally, PC1 for each will be discussed, since only the first component will be used to construct the subindices. PCA will then be conducted on the subindices to compute the COVID Infection Index (CII).

	SOCIOECON				**HLTHAGE**				**HSELIV**		
Comp.	*ESS WRKER*	*RENT 35%*	*NO INSUR*	*(%)*	*OVER 65*	*HIGH CHOL*	*DISABL*	*(%)*	*AVG HH*	*GROUP Q*	*(%)*
PC1	0.628	0.518	0.581	60.94	-0.568	-0.631	-0.528	62.420	0.707	-0.707	56.70
PC2	-0.155	0.814	-0.559	24.02	0.624	0.089	-0.777	23.670	0.707	0.707	43.30
PC3	-0.763	0.261	0.592	15.04	0.537	-0.770	0.343	13.910			

Table 5. The individual PCA loadings and percent of variance explained component for each component and subindex. PCA was conducted at the census tract level (n=2105).

PCA was conducted on the standardized variables of each subtheme. According to table 5, nearly 60% of the variance in SOCIOECON and HLTH AGE is captured by the PC1. For HSELIV, 57% is captured by the PC1.

For the SOCIOECON subindex, ESS WRKER, RENT 35%, and NO INSUR have similar positive loadings on the PC1, suggesting that they vary together. Many ESS WRKER are underpaid and possibly struggling with household expenses. Some of these jobs include people who work in Health care support positions, home health aides, and cleaning positions.[52] And compared to non-essential workers, essential worked have higher uninsured rates. [53]

As for the HLTHAGE subindex, OVER 65, HIGH CHOL, and DISABL have similar negative loadings on PC1. The less people over the age of 65, the less likely we are to see high cholesterol and disability rates. This is reasonable given that the senior population sees higher rates of high cholesterol and develop one of the six difficulties listed by CENSUS over time.

Lastly, the HSELIV subindex loadings reveals an issue with the data. Many census tracts have GROUP Q equal to 0, results in factors (or scores) for some tracts being nearly identical to AVG HH. This results in perpendicular loadings, where AVG HH loads (.707,.707) on (PC1, PC2) which is equivalent to $\left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$. From a geometric perspective, this is equivalent to a 45-degree angle between PC1 and PC2 in the upper right quadrant. Similarly, the vector loadings of GROUP Q are 45-degree angle in the upper left quadrant—see figure 6. And so, the AVG HH standardized variable was decided to represent the HSELIV subindex alone.

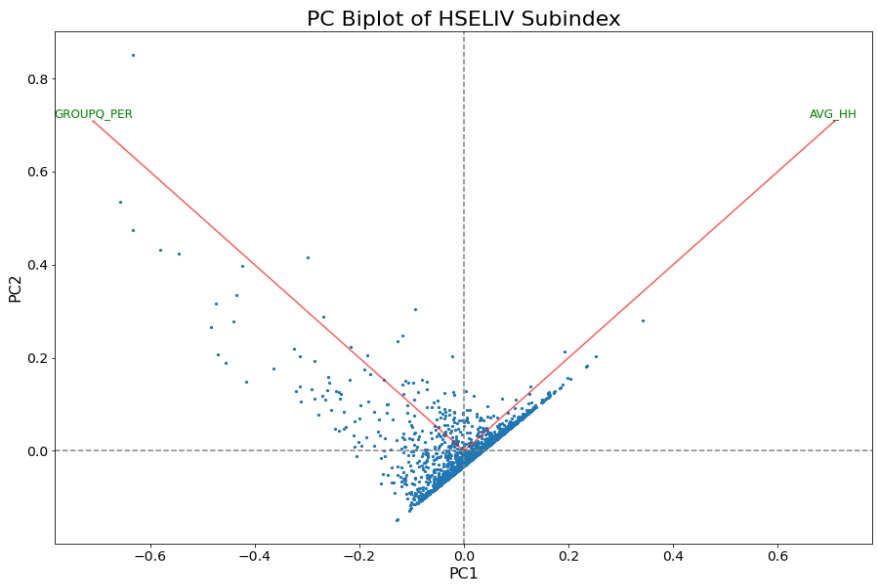


Figure 6. PC biplot of AVG HH AND GROUP Q

After compiling all the necessary subindices to compute the CII, PCA was run on the subindices. From table 6, we see that SOCIOECON and HSELIV loads highly on PC1, indicating that the lower the socioeconomic status of a person, the higher chances they have living in a more crowded household. HLTHAGE load less on PC1 and highly on PC2, suggesting that health and age, as defined here, does not have vary as much when SOCIOECON and HSELIV do.

	CII			
Comp.	*SOCIOECON*	*HSELIV*	*HLTHAGE*	*(%)*
PC1	0.703	0.71	0.040	51.00
PC2	0.173	-0.117	-0.978	34.12
PC3	-0.69	0.695	-0.205	14.88

Table 6. Final PCA loadings and percent of variance of CII.

Using PC1, we computed the CII and mapped the results to visually inspect its performance. After CII was computed, it was rescaled to the range of 0 to 1 using the min-max transformation. Jenks classification with five breaks was used to visualize CII and INF_RATE.

In figure 7, the map of CII on the left follows the general trends of INF_RATE (on the right). Cold zones like Steinway and Brooklyn heights were captured by the CII. In lower, mid, and upper Manhattan some of the lowest CII scores were seen, ranging from 0.00 to 0.22, along with the lowest infection rates. Hot zones in like in south-west Bronx, Jackson Heights and flushing in Queens, and Bay Ridge, Sunset Park, and Borough Park in Brooklyn were captured by the CII.

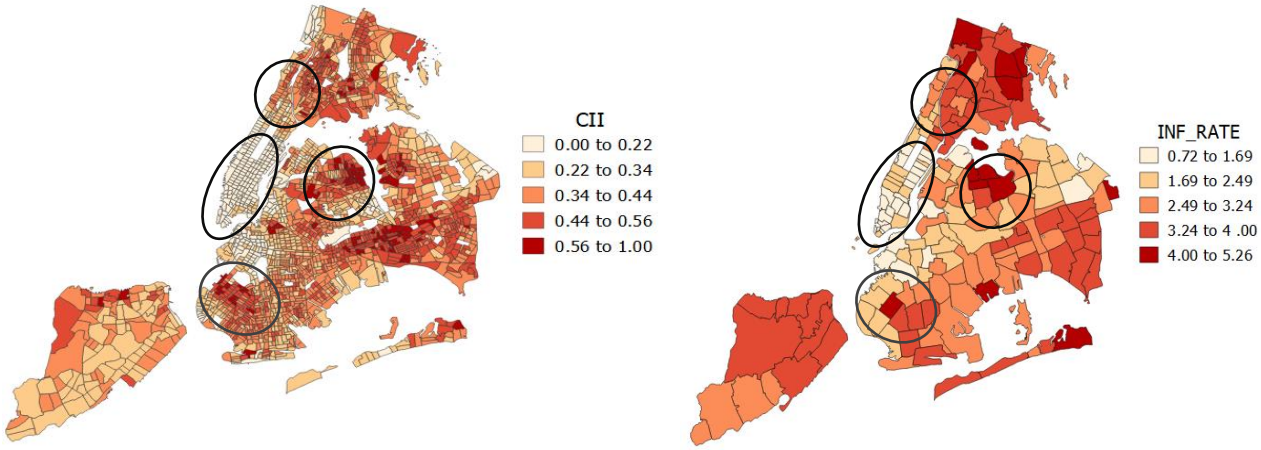


Figure 7. CII map to the left and INF_RATE to the right.

GWPCA

Like GWR, GWPCA was conducted using an adaptive gaussian kernel to ensure that each local PCA analysis contains enough samples. Variables were standardized prior to running the analysis. To be consistent, AVG HH was used. We will only discuss the geographically weighted (GW) PC1 for each sub index calculation.

From figure 9, we see that ESS WRKER, RENT 35%, and NO INSUR load negatively on GW PC1 in wealthier areas like Fresh Meadows and College Point in Queens and Throgs Neck and Pelham Bay in the Bronx. In these areas, these three variables are highly negatively related meaning that if one decreases, the other two will as well. Areas beyond these neighborhoods, the three variables load positively on GW PC1. Specifically, NO INSUR and RENT 35% loads high in West Queens while NO INSUR and ESSWRKER load similarly in Manhattan and areas closest Manhattan.

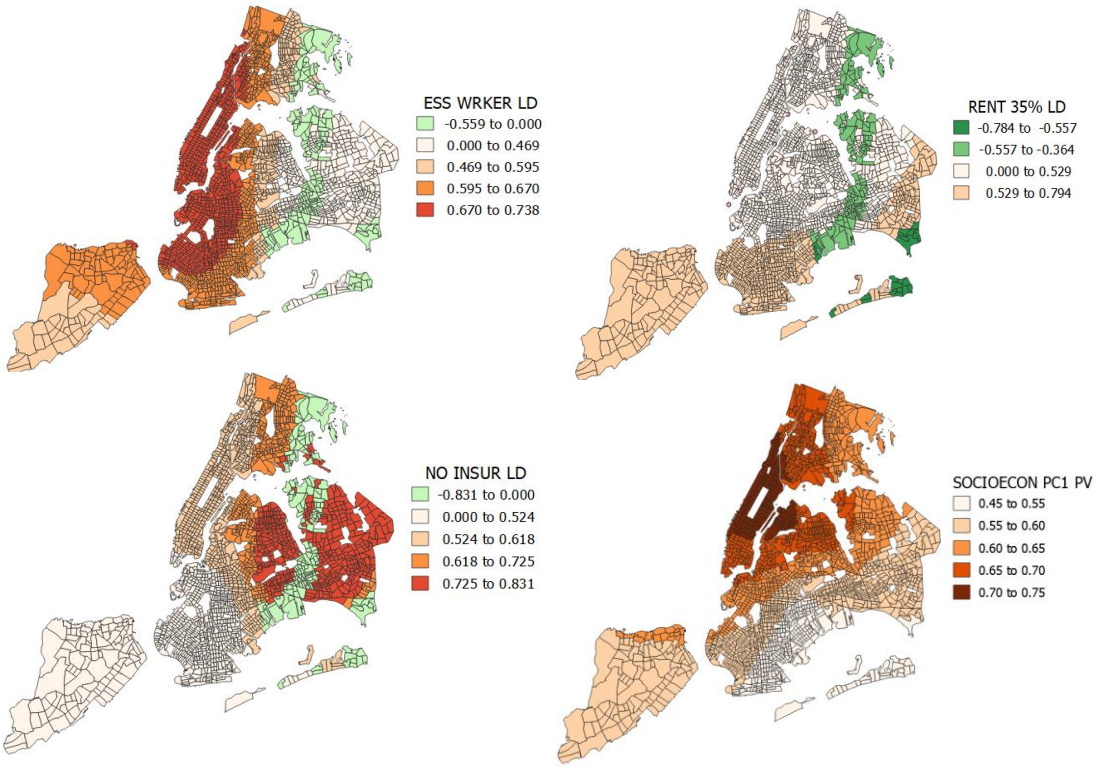


Figure 8. ESS WRKER, RENT 35%, and NO INSUR loadings on the GW PC1 of the CII PCA analysis. The bottom right map displays the percentage of variance explained of the GW PC1.

GW PC1 of the SOCIOECON sub index explains nearly 70% to 75% of variance in Mid and Upper Manhattan. The further away from Manhattan, the less variance the GW PC1 explains, seeing numbers as low as 45% to 55% of explained variance in areas like Canarsie, Sheepshead Bay and Rockaway Park.

Compared to the loadings from the non-GW PCA analysis, Figure 9 demonstrates a similar relationship among the variables HLTHAGE variables OVER 65, HIGH CHOL, and DISABL. Each variable loads negatively on the HLTHAGE GW PC1 across NYC. The variables load similarly in East Bronx, West Queens, Brooklyn, suggesting that the GWPCA analysis captured a relatively stable relationship among OVER 65, HIGH CHOL, and DISABL in these areas. As any one of the variables decreases, the other two decrease as well.

In contrast, each variable displays a high negative loading in specific areas of NYC, while the remaining load is much less in these areas. In Manhattan and West Bronx, HIGH CHOL alone demonstrates high negative loadings on HLTHAGE GW PC1. High negative loadings are found for OVER 65 alone in East Queens. As for DISABL, high negative loadings are found in South Staten.

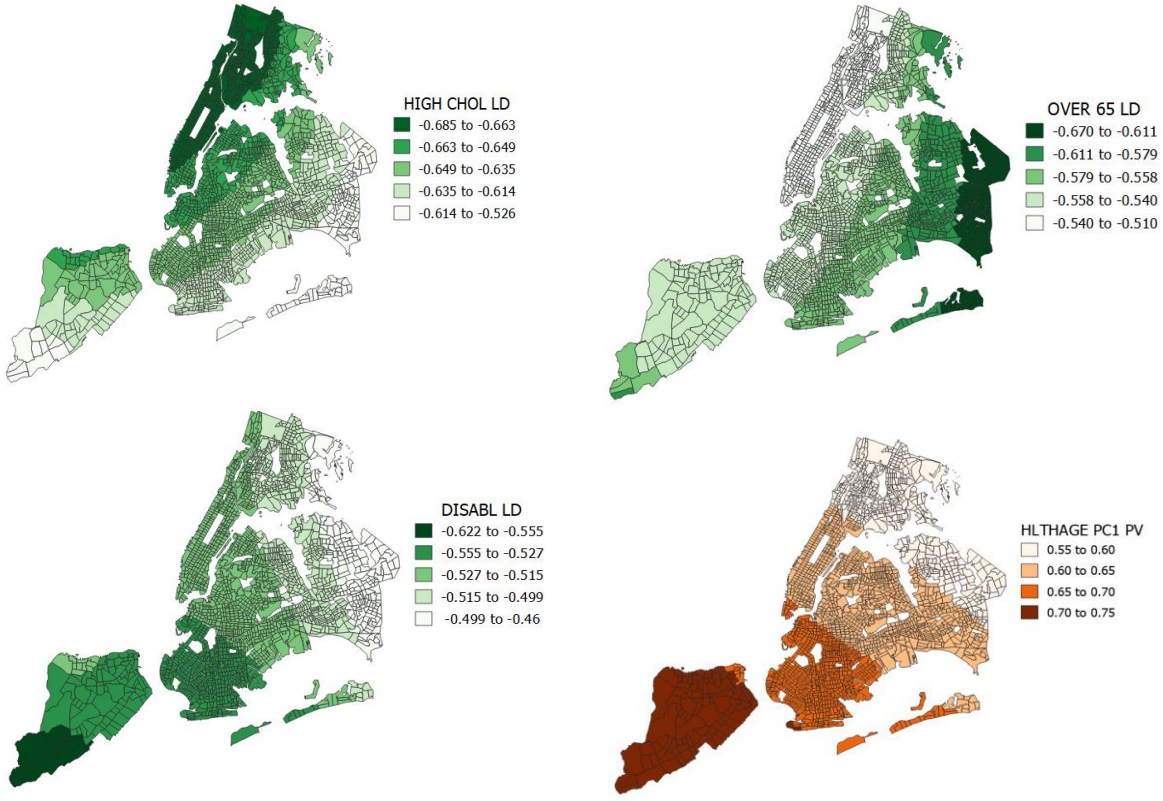


Figure 9. HIGH CHOL, OVER 65, and DISABL loadings on the GW PC1 of the CII PCA analysis. The bottom right map displays the percentage of variance explained of the GW PC1.

GW PC1 of the HLTHAGE sub index explains majority of variance in Staten Island and Brooklyn, ranging from 65% and 75% OVER 65 alone in East Queens. The percentage of variance explained decreases in Manhattan, Queens, and the Bronx, ranging from 55% to 65%.

From figure 10, we see that HSELIV, HLTHAGE, and SOCIOECON for the most part load differently on GW PC1 of the CII. This suggests that the subindices possess inverse relationships—i.e., one subindex increases while the other decreases. GW PC1 of the CII does not account for majority of the variance among the three sub-indices, with the maximum variance captured being between 50% to 55%

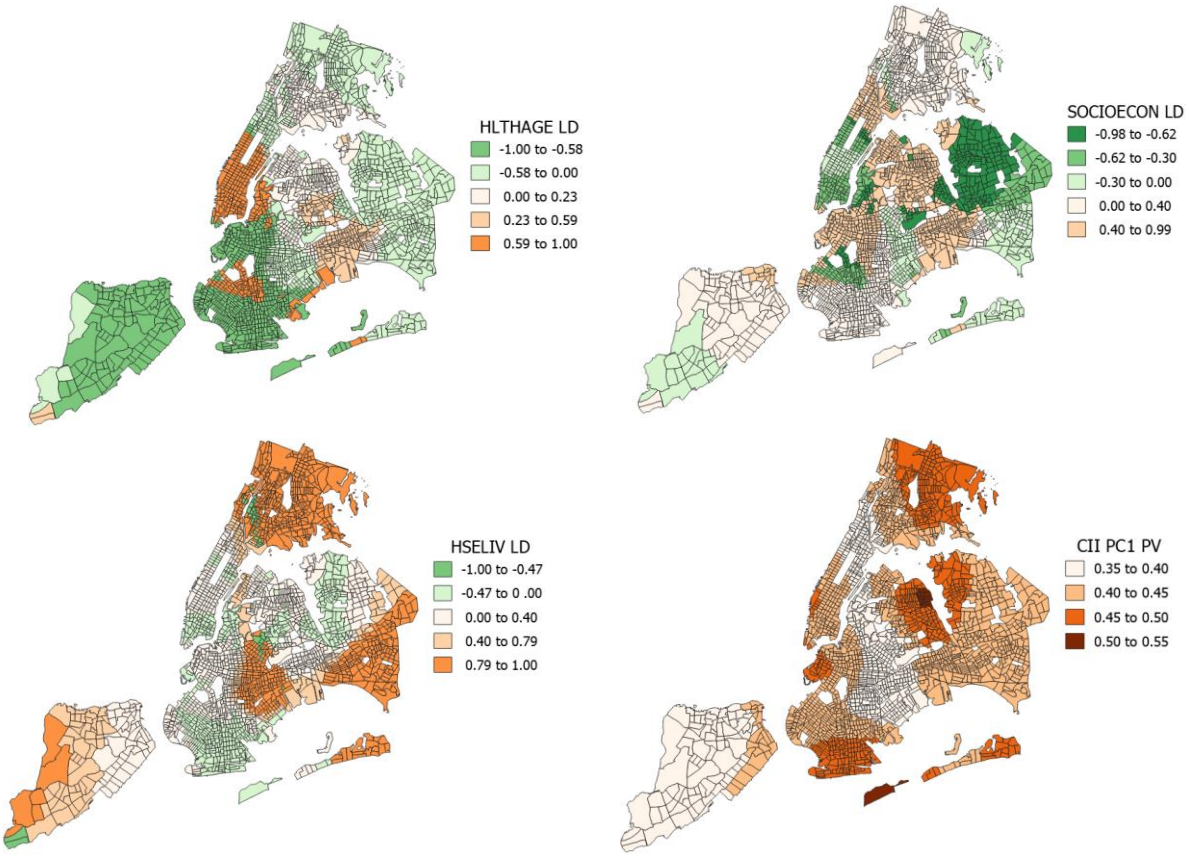


Figure 10. HLTHAGE, HSELIV, and SOCIOECON loadings on the GW PC1 of the CII PCA analysis. The bottom right map displays the percentage of variance explained of the GW PC1.

According to figure 11, we see that the GW CII does not capture much of the INF_RATE patterns or distribution. The distribution of GW CII is heterogeneous, meaning that the values are scattered and do not seem to accurately capture the ground truth of COVID infection. In the next section, we will discuss some possibilities as to why this is.

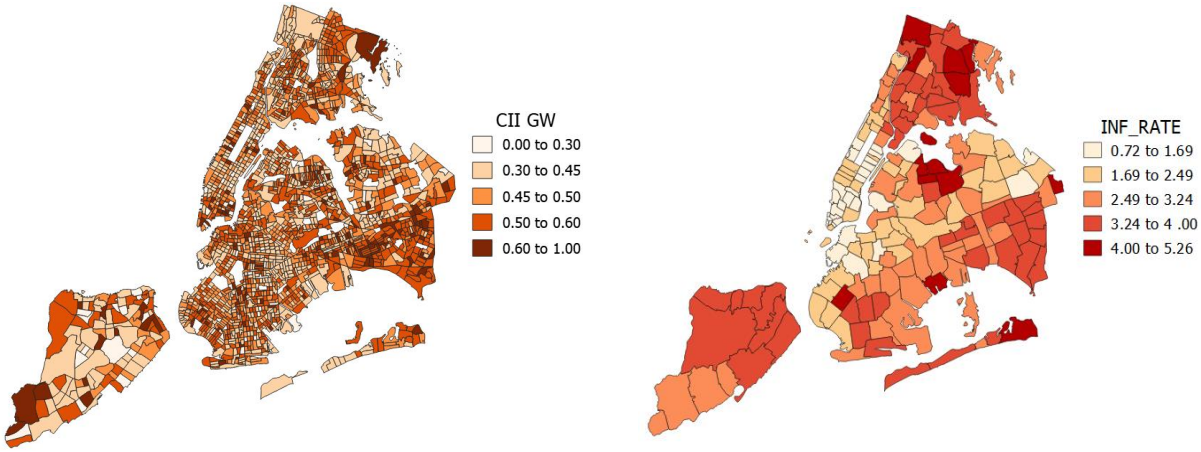


Figure 11. Geographically Weighted CII map to the left and INF_RATE to the right.

Discussion

Ultimately, facing COVID takes collective effort. Everyday decisions of everyone should follow some line of precautions and rationale. Should I attend a small get together with friends? do I really need to go to a bar? should I take a vacation this year are all questions with obvious answers to some of us. Unfortunately, not all people are equipped or capable to make good decisions, a trait that has been linked to poverty, trauma, and stress, inter alia. [54] [55] This disparity is difficult to capture with data alone, given the mechanisms behind deciding what is best for oneself. Such a disparity is contingent on (but not limited to) cultural, social, neurophysiological, educational, socioeconomical, and ethical factors. Many youths, for example, are more willing to disobey COVID guidelines or congregate because the virus will not directly affect them. Others may view their freedom to socialize and congregate as more valuable than keeping the public safer, while others may not feel this way but cannot regulate the impulse to socialize—and so on.

Despite all that was not captured, this study has captured some measurable contributions to COVID infection. In addition, this study has reiterated the importance of considering the spatial properties of COVID—the infection varies spatially. It also seems like we have captured a composite way of assessing COVID infection, prior to the spread. The CII has followed some of the visual patterns of COVID infection. However, the CII was not validated statistically and working towards a way of doing this is important to assessing the reliability of the index. More importantly, the GW counterpart of the CII did not perform as anticipated. This does not undermine the quality and usage of the techniques and results presented here. There are many possible reasons as to why GWPCA did not perform as well as PCA.

Any GW model is complex since it is largely dependent on the type of kernel selected (adaptive or fixed), the spatial weighing function (gaussian, bi-squared, exponential, etc), and the distance metric. The kernel type is often easy to decide upon: a fixed kernel performs better with large and dense samples whereas an adaptive kernel performs better with sparse and scattered samples. To assess the reliability of GWR and GWPCA here in this study, rigorous work needs to be done in determining which spatial function is optimal and which distance metric accurately captures the spatial relationship among the samples. Testing different spatial functions and comparing the adjusted Akaike Information Criterion (AICc) of each could possibly serve as one approach.

As for the distance metric, Euclidean distance was used here. The Euclidean distance may have inaccurately represented the distance between one census tract to another. Census tracts are connected by roadways, bridges, and highways. A census tract in say Staten Island may be close to a Census tract in Brooklyn when in fact they are further

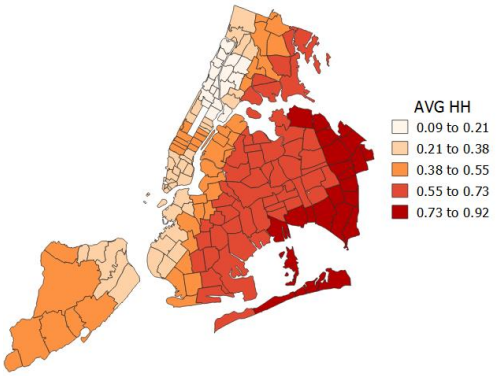
away if we use non-Euclidean distance metrics such as road network distances or travel time metrics.[56]

In addition to considering the type of spatial weight function and distance metric, modifying the GW CII calculation may also be helpful. In non-GW CII, we considered the scores of PC1 as measures of each subindex. We considered this largely because PC1 explained most of the variance in the variables used. However, as we have seen, this is not always the case for GWPCA: at some census tracts GW PC1 accounted for majority of the variance, but for others, GW PC1 possibly accounted for the least amount of variance. It may be worthwhile using scores of different PC for each census tract, depending on whether the PC selected accounts for the most variance. Approaching GWPCA with these considerations may improve the performance of GW CII and this technique in further studies.

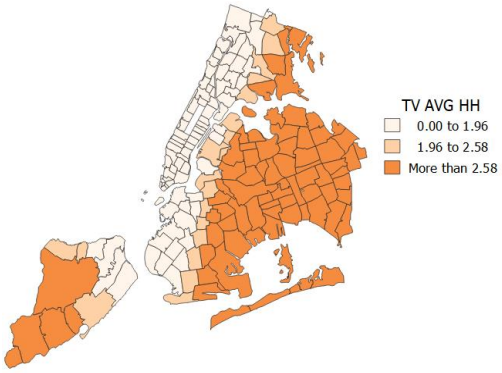
APPENDIX A:

AVG HH

Local Coefficients

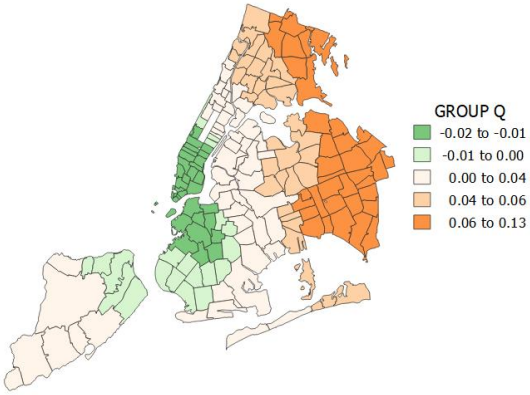


Local t-values

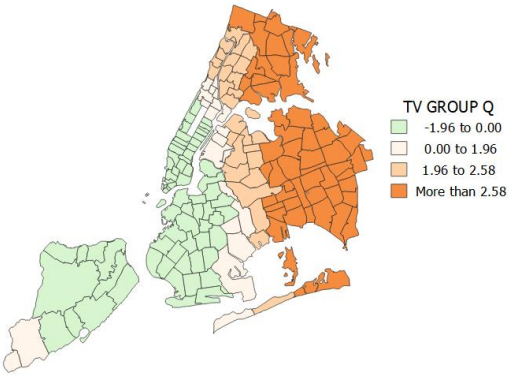


GROUP Q

Local Coefficients

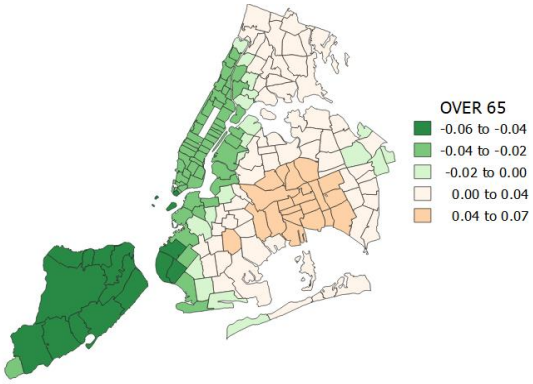


Local t-values

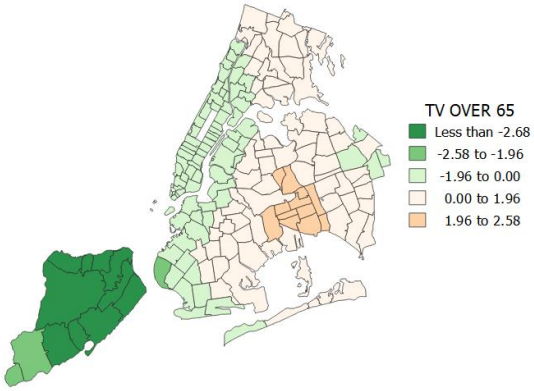


OVER 65

Local Coefficients

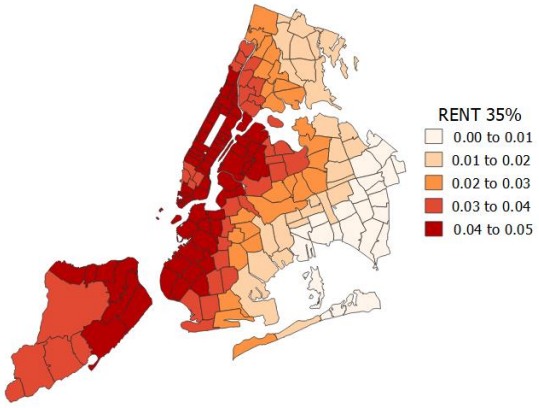


Local t-values

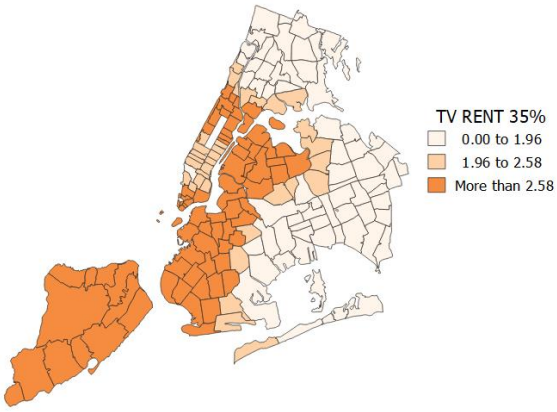


RENT 35%

Local Coefficients

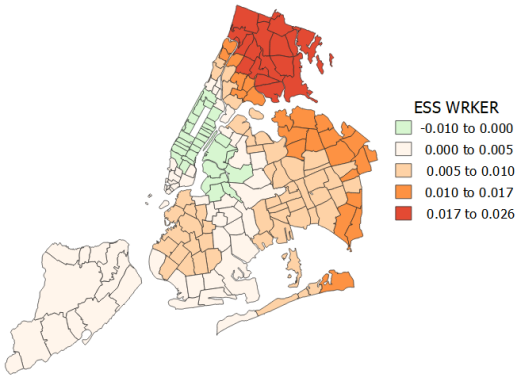


Local t-values

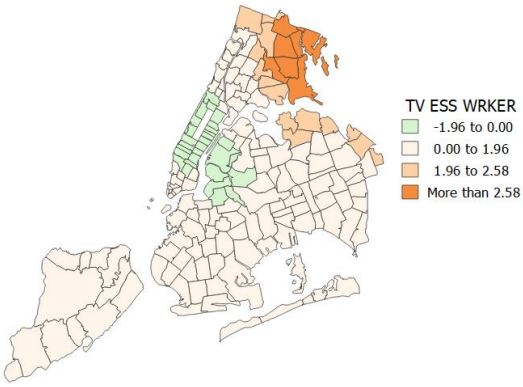


ESS WRKERS

Local Coefficients

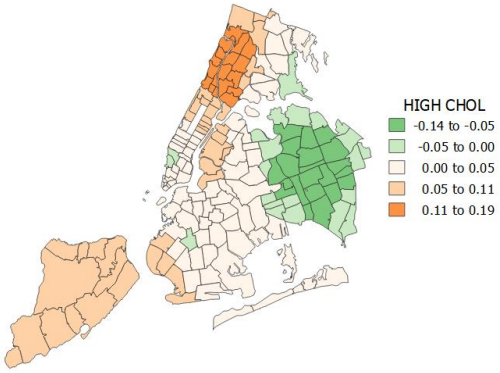


Local t-values

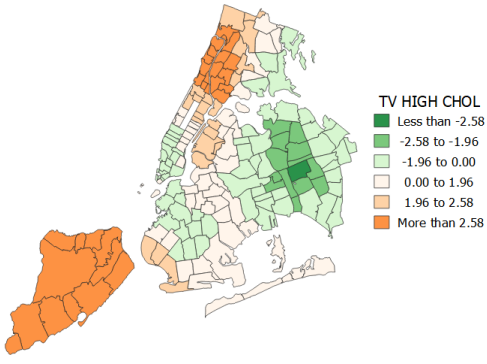


HIGH CHOL

Local Coefficients



Local t-values



[1]https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted

[2]https://www.ilr.cornell.edu/work-and-coronavirus/student-voices/what-rights-do-essential-workers-have-nys-and-nyc

[3]https://www.governor.ny.gov/sites/governor.ny.gov/files/atoms/files/EO202.6.pdf

[4]https://www.nytimes.com/2020/08/02/nyregion/nyc-subway-coronavirus-safety.html

[5]https://static1.squarespace.com/static/5bc63eb90b77bd20c50c516c/t/5f74915264a865029dafa27c/1601474930418/APTA+Covid+Best+Practices+-+09.29.2020_update.pdf

[6] https://gothamist.com/news/new-study-finds-no-direct-link-between-subway-covid-19-spread

[7] https://www.cdc.gov/coronavirus/2019-ncov/daily-life-coping/using-transportation.html

[8] https://creakyjoints.org/living-with-arthritis/coronavirus/daily-living/is-it-safe-carpool-high-risk-covid-19/

[9] https://www.nyu.edu/about/news-publications/news/2020/july/covid-safety-training-rideshare.html

[10] https://www.cuimc.columbia.edu/news/crowded-homes-poor-neighborhoods-linked-covid-19

[11] https://jamanetwork.com/journals/jama/fullarticle/2767631

[12] Office of the new york city comptroller. (2018, Oct 11,). *Tenders & Projects Information* https://www.emis.com/php/search/doc?pc=US&dcid=629790039&primo=1

[13] Steven Ruggles, & Susan Brower. (2003). Measurement of household and family composition in the united states, 1850-2000. *Population and Development Review, 29*(1), 73-101. 10.1111/j.1728-4457.2003.00073.x

[14] https://www.census.gov/2018censustest/gq

[15] Jiménez, M. C., Cowger, T. L., Simon, L. E., Behn, M., Cassarino, N., & Bassett, M. T. (2020). Epidemiology of COVID-19 among incarcerated individuals and staff in massachusetts jails and prisons. *JAMA Network Open, 3*(8), e2018851. 10.1001/jamanetworkopen.2020.18851

[16] https://www.nytimes.com/interactive/2020/us/covid-college-cases-tracker.html

[17] https://www.health.ny.gov/press/releases/2020/docs/nh_factors_report.pdf

[18] https://www.apa.org/pi/ses/resources/publications/task-force-2006.pdf

[19] Christopher R. Huber, & Nathan R. Kuncel. (2016). Does college teach critical thinking? A meta-analysis. *Review of Educational Research, 86*(2), 431-468. 10.3102/0034654315605917

[20] https://societyhealth.vcu.edu/work/the-projects/why-education-matters-to-health-exploring-the-causes.html

[21] https://cs.stanford.edu/people/eroberts/cs181/projects/digital-divide/start.html

[22] https://www.healthaffairs.org/doi/full/10.1377/hlthaff.21.2.60).

[23] https://www.nejm.org/doi/full/10.1056/NEJMp2008017

[24] https://www.feedingamerica.org/sites/default/files/2020-10/Brief_Local%20Impact_10.2020_0.pdf

[25] https://www.ceinsys.com/blog/point-of-interest-really-necessary-for-mapping/).

[26] https://www.w3.org/2010/POI/wiki/Terminology

[27] https://www.npr.org/2020/11/19/936490226/some-faith-leaders-defiant-others-transparent-over-covid-19-outbreaks

[28] https://www.nytimes.com/2020/07/08/us/coronavirus-churches-outbreaks.html

[29] https://www.jewishvirtuallibrary.org/what-is-shabbat-jewish-sabbath

[30] https://www.ucg.org/bible-study-tools/bible-questions-and-answers/which-day-is-the-sabbath-according-to-the-bible

[31] https://pluralism.org/jum%E2%80%99ah-the-friday-prayer

[32] https://www.nbcnews.com/news/nbcblk/black-barber-shops-salons-safe-havens-cultural-chats-n1184691

[33] case_study_barbershop.pdf (harvard.edu)

[34] https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2770975

[35] https://www.supermarketnews.com/online-retail/nearly-80-us-consumers-shopped-online-groceries-covid-19-outbreak

[36] https://www1.nyc.gov/site/hra/help/snap-online-shopping.page

[37] https://www.nytimes.com/wirecutter/blog/laundromats-during-coronavirus/).

[38] https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/evidence-table.html

[39] https://www.cdc.gov/mmwr/volumes/69/wr/mm6936a3.htm

[40] https://www.lhsfna.org/index.cfm/lifelines/august-2020/are-construction-workers-at-higher-risk-for-covid-19-complications/#:~:text=After%20taking%20age%2C%20medical%20conditions,of%20workers%20across%20all%20industries

[41] https://www.kff.org/medicare/issue-brief/one-in-four-older-adults-report-anxiety-or-depression-amid-the-covid-19-pandemic/

[42] https://labblog.uofmhealth.org/rounds/loneliness-doubled-for-older-adults-first-months-of-covid-19

[43] https://khn.org/news/seniors-in-low-income-housing-live-in-fear-of-covid-infection/,

[44] https://www.brookings.edu/blog/the-avenue/2020/03/16/for-millions-of-low-income-seniors-coronavirus-is-a-food-security-issue/

[45] https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1003379,

[46] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7226715/

[47] https://www2.census.gov/geo/pdfs/reference/glossry2.pdf).

[48] https://www.npr.org/sections/thetwo-way/2013/07/01/197623129/the-zip-code-turns-50-today-here-are-9-that-stand-out#:~:text=The%20Most%20Spacious%3A%2089049%20(map,than%20the%20state%20of%20Maryland)

[49] https://www2.census.gov/programs-surveys/acs/tech_docs/subject_definitions/2019_ACSSubjectDefinitions.pdf

[50] (Guo HJ, Sapra A. Instrumental Activity of Daily Living. [Updated 2020 Mar 18]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK553126/ ,

[51] Edemekong PF, Bomgaars DL, Sukumaran S, et al. Activities of Daily Living. [Updated 2020 Jun 26]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK470404/

[52] https://www.brookings.edu/research/essential-but-undervalued-millions-of-health-care-workers-arent-getting-the-pay-or-respect-they-deserve-in-the-covid-19-pandemic/

[53] https://www.kff.org/policy-watch/taking-stock-of-essential-workers/.

[54] https://digitalcommons.buffalostate.edu/cgi/viewcontent.cgi?article=1001&context=pubassistdevprograms

[55] https://www.jrf.org.uk/report/how-poverty-affects-peoples-decision-making-processes).

[56] Binbin Lu, Martin Charlton, Paul Harris & A. Stewart Fotheringham (2014) Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data, International Journal of Geographical Information Science, 28:4, 660-681, DOI: 10.1080/13658816.2013.865739