

DEMOGRAPHIC INCOME VERSUS GEOGRAPHICAL TRENDS OF COVID-19: DATA ANALYSIS REPORT

Chris Chiang
PSTAT 120C
Professor Michael Gu

ABSTRACT

COVID-19 trends do not affect individual populations, and especially not in specific regions in the United States in the same way. Demographic factors, such as population density directly affect county-sorted infection rates, deaths, and positivity. However, these factors have only grazed the many reasons for the disparity of infection rates and deaths among individuals. In this data report, income levels (sorted by county within the US per state) will be considered and compared with previous plots and graphs. This report mainly focuses on California, although other states along with US charts are also mentioned and reported on.

Data from the USDA data reports regarding median household income in 2019 for each county in the US (also per state) will be used. The dataset also includes unemployment rates from 2011-2019, but I will only look at median household income levels in 2019. Although 2019 was only a year of the virus' first discovery and not a year of rapid COVID-19 infection, 2020 income has not been reported so far. Income levels follow a similar rate of increase per year, at least in a given state; therefore, data from any year provides a reasonably equivalent comparison between multiple regions. I want to figure out the relationship between county-level income data and coronavirus infections (infections and death count) using tools like 7-day averages and state-county maps as previously noted. It is important to understand this relationship, as it is crucial to accurately predict future potential infection/deaths as well as the urgency of vaccine distribution in specific populations.

INTRODUCTION and METHODOLOGY

Firstly, I will implement my dataset (UnemploymentReport.csv) from USDA which I can turn into a CSV file for R, in order to create models and plots regarding county household income level, which could be used by state as well as for the entire US as a whole. There will be numerical and interpretive questions I have created, and various graphs accompanying each question. Questions will each have an introduction, explaining why it is being asked and its relevance to the central purpose, and then ask for an analysis of the plots created to accurately answer the question. To thoroughly understand the new plots for income vs older DA2 graphs, I will include topics such as: infection during peak periods, the overall infection/death count, how income affects positive rates for a particular state, and measures to reduce potential disparities reported.

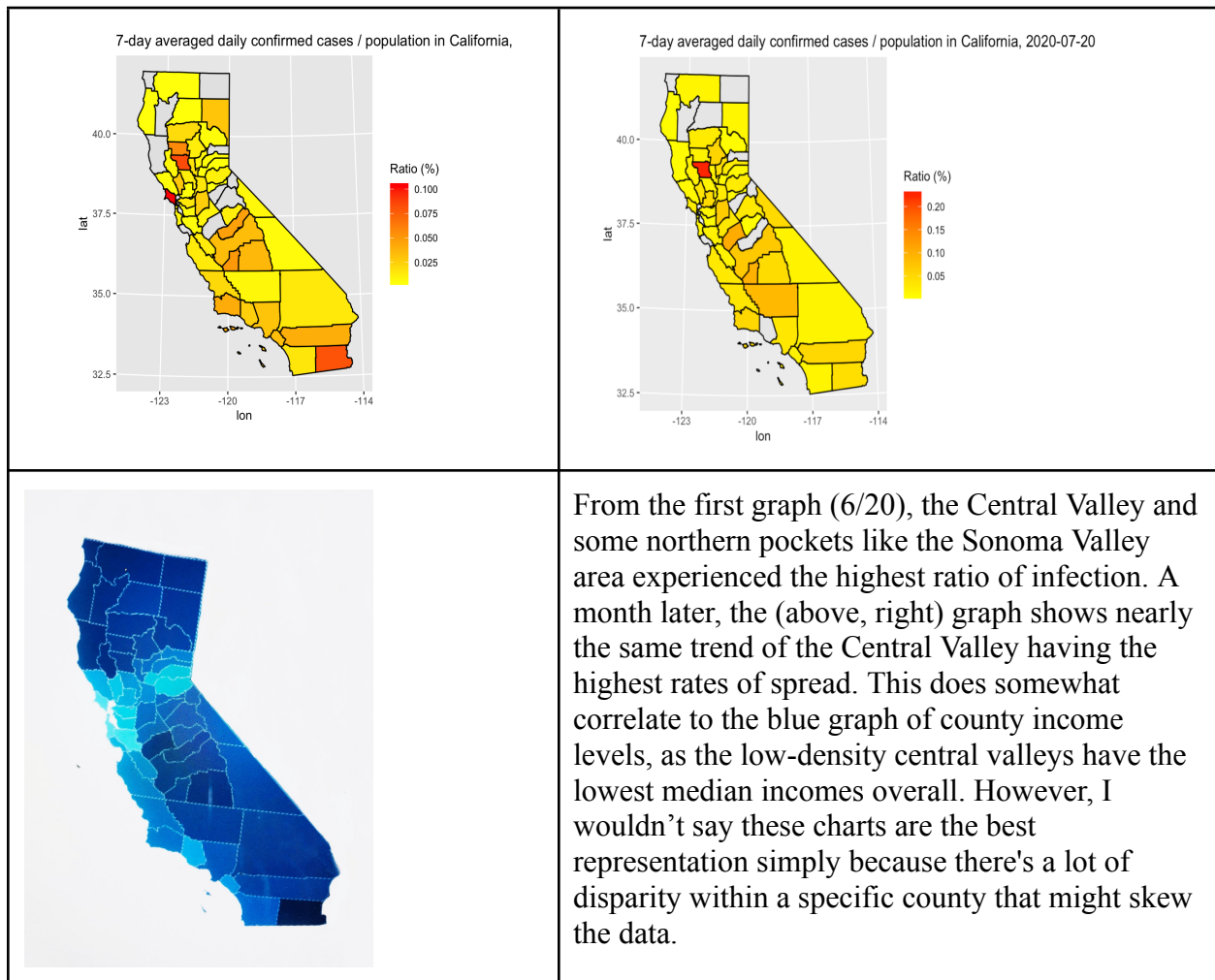
Secondly, I will look at 2 research articles. The first one is titled: “*Disparities In Outcomes Among COVID-19 Patients In A Large Health Care System In California*” which was published under the Health Affairs journal on 21 May 2020. The article was written by many authors: Kristen M. J. Azar, Zijun Shen, Robert J. Romanelli, Stephen H. Lockhart, Kelly Smits, Sarah Robinson, Stephanie Brown, and Alice R. Pressman. The second titled “*GIS-based spatial modeling of COVID-19 incidence rate in the continental United States*” is written by Abolfazl Mollalo, Behzad Vahedi and Kiara Rivera published under the Science of the Total Environment journal on 1 August 2020.

In my conclusion of the project, I will then compare my questions/plots and then analyze the trends that were discovered and studied in my research papers. Using these sources, I will attempt to thoroughly answer, reflect, and explain my research topic.

REPORT ANALYSIS QUESTIONS:

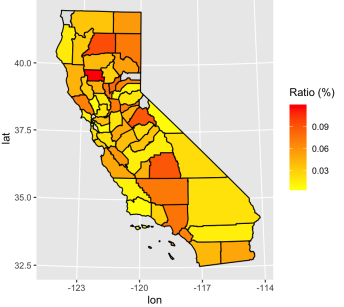
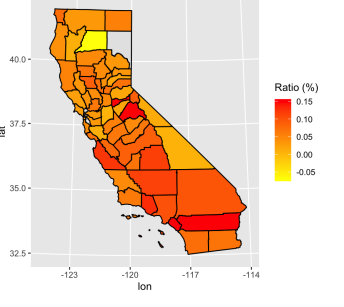
RESULTS AND INTERPRETATION

1. During the peak surges in California, is there a higher disparity of infection rates/cases due to income? Make maps for the 7-day average of daily county-level confirmed cases over the county population in California on June 20, July 20. Compare that to the income levels of each county/region. Is there a noticeable relationship, explain why there is or isn't?

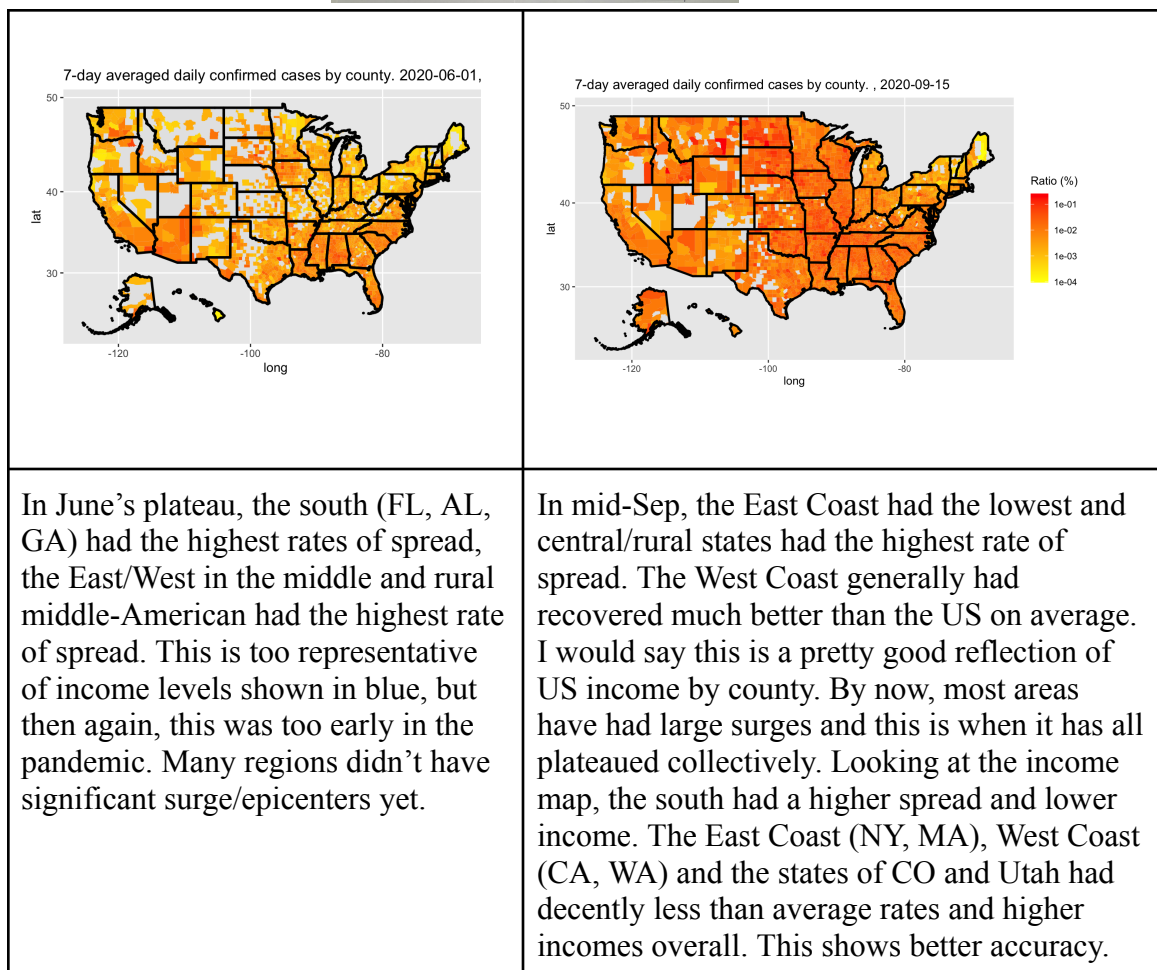
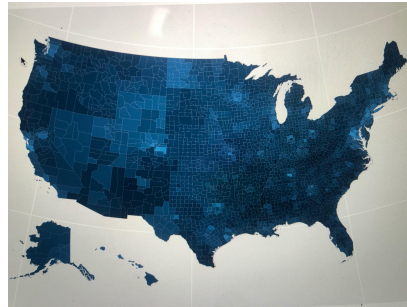


2. During the holiday months, there was an exponential surge of spread in California and the West Coast in general. Answer the first question, instead, with the dates of Dec 1 and the end of this month, on Dec 30.

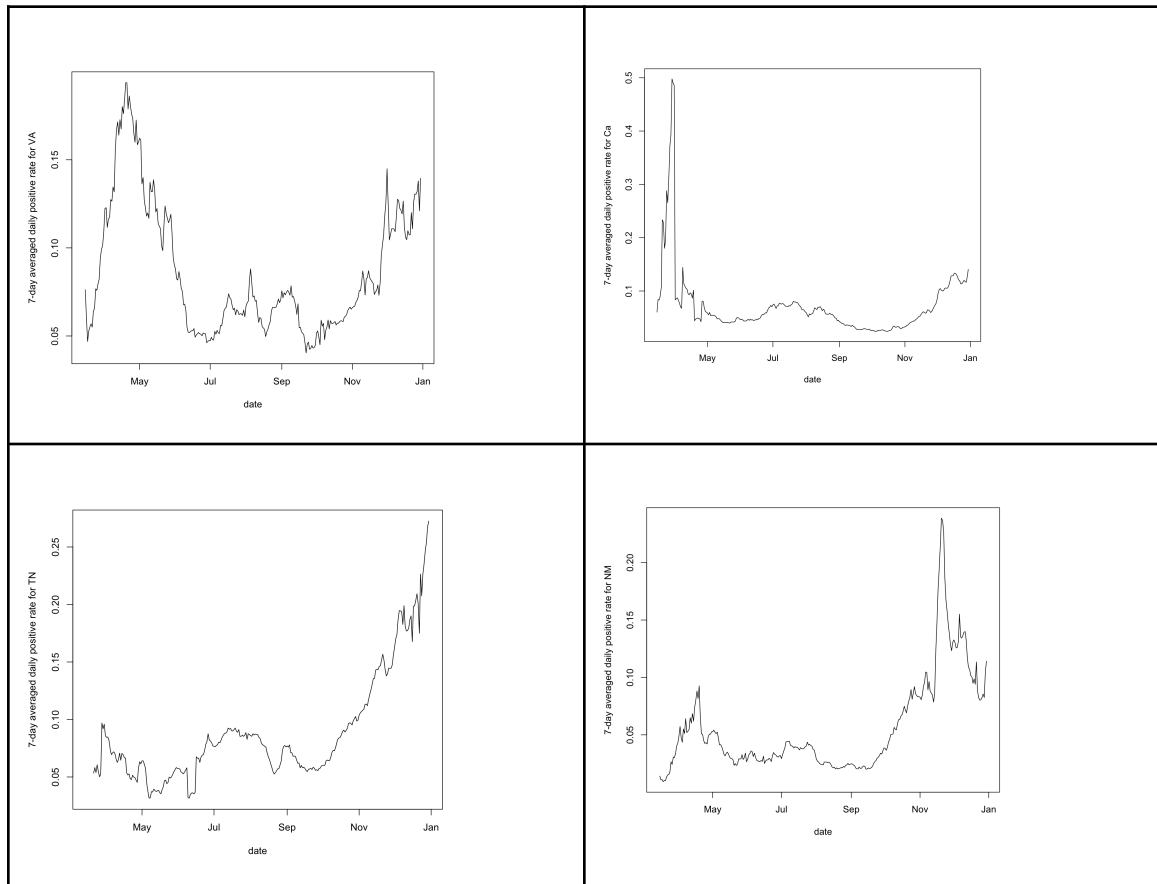
**referring to blue graph for income levels in Question 1

 <p>7-day averaged daily confirmed cases / population in California, 2020-12-01</p>	 <p>7-day averaged daily confirmed cases / population in California, 2020-12-30</p>
<p>On December 1, I can see that there was rapid spread but it was not very consistent. There are pockets in SoCal (LA and Kern), up in the Central and North valleys. The Bay Area, with the highest median incomes experienced rates similar to the state as a whole and while rural lower income areas did have a higher rate, the disparity doesn't seem to have an extremely casual relationship during this time.</p>	<p>By the 30st, the Bay Area seemed to be closer to the state average than ever before, although with much higher income levels. The spread centered in Los Angeles, Ventura and Orange County (SoCal in general). These counties have a slightly above average median income, yet have the highest rates of spread. The conclusion from these 2 problems is that the OVERALL state/county income median is not the brightest predictor of the rate of spread. Other income data sorted differently may be more resourceful.</p>

3. The country experienced a plateau in cases in the early Summer, with the lowest number of daily cases (hovering at about 20,000) since the pandemic started. After the second surge, cases plateaued surrounding mid-September. Compare the 7-day average rate per county for the US for June 01 and Sep 15 and the income levels of the US. Is there a trend? Explain.



4. Compare overall trending positive rates (from the period of March 15 to December 31) for some higher-density, higher-income states (Virginia and CA), then a few higher-density but low-income states (Alabama and SC). What patterns are seen? Compare VA vs SC and CA vs NM since they are respectively in a similar geographic area.



Looking at the overall state level, there is no obvious relationship. Comparing VA and TN, the states had different outbreaks at different times (VA during late Spring). During the winter outbreaks, the same pattern of exponential increase is seen, though TN (lower income) is slightly higher.

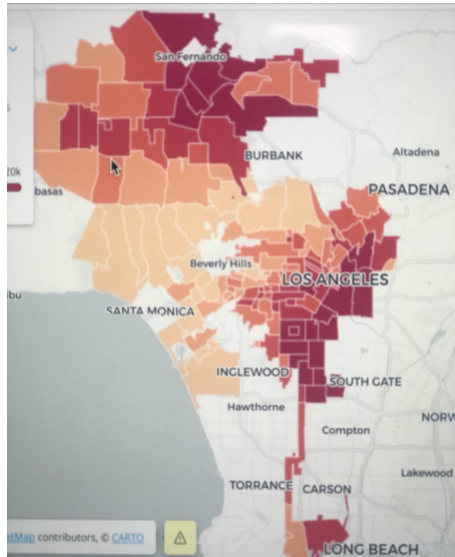
Comparing CA vs NM, they both generally stayed at 5%-10% positivity rates throughout, except for small surges that went back down quickly. Generally, during the summer surge, both states increased a little bit and then stayed at under 5% during early fall/late summer.

5. *Is an individual county's median income the best predictor of rate of spread ? If yes, explain why and if not, what would be a good alternative variable to use that still considers income levels?*

I don't believe county income is the most effective because there is a lot of variation within the county limits, and a large population of different types/classes of people already. Income itself is an arguably good measure but comparing the counties vs overall state isn't the most resourceful. Using smaller areas like cities or neighborhoods and then comparing them to the overall county or even the overall state would be a little bit more accurate. There is less variation and difference in just a small city/neighborhood with only a few thousand people versus a county with millions. A statistic with a large number of samples (in this case, reports) that is summarized into one number (income) doesn't account for significant and predictive disparities within.

6. Since median income data collected from an entire county (in comparison to the respective state) is not especially accurate, analyze neighborhoods within a certain county and compare their infection rates vs income. Would this be a better predictor? You may consult the internet for available maps.

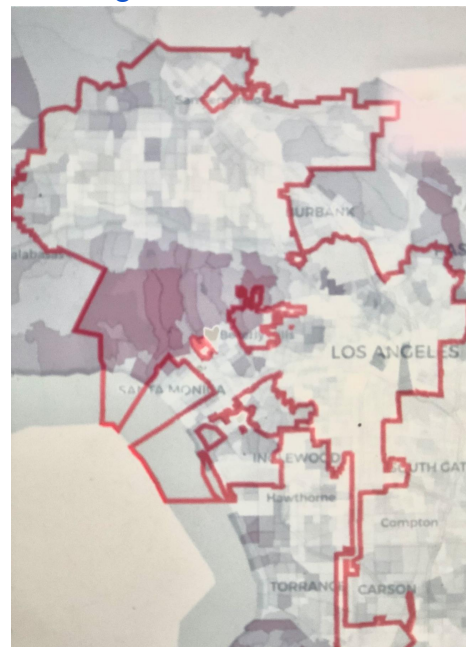
<https://www.nbclosangeles.com/news/coronavirus/southern-california-coronavirus/los-angeles-la-neighborhood-coronavirus-cases-deaths-map-covid-19/2490987/>



This data is the 14-day rate in 12/21

Darker= lower rate
Lighter= higher rate

<https://www.city-data.com/income/income-Los-Angeles-California.html>



Median Household Income in 2019

Darker= higher income
Lighter= lower income

I used data from these 2 links to compare LA neighborhood 14-day average rates (Dec 21) and then, income. Note that the color legend in both charts work inversely. I can see that there is a very noticeable pattern between the two factors. They correlate almost perfectly with the highest spread in low-income neighborhoods and vice versa.

From the first chart,

1. HIGHEST spread- San Fernando Valley (especially North and NE), South and East Los Angeles
2. MIDDLE- Southern SF Valley, Central and North LA.
3. LOWEST spread- The Westside, Hills, and Coastal regions like Santa Monica/LAX area

From the 2nd income chart, I can observe the same trend

1. LOWEST income- South and East LA, North and East San Fernando Valley
2. MIDDLE- Central LA, Central SF Valley and pockets within North LA/Verdugo Mountains
3. HIGHEST income- Beverly Hills/Crest, Bel-Air, Brentwood all of which are all wealthy, hillside neighborhoods in the Westside.

7. How do you think income affects covid-19 rates/spread/positivity and its cumulative/daily total cases and deaths? Why is this the case, and what other demographic factors can income level represent or indicate?

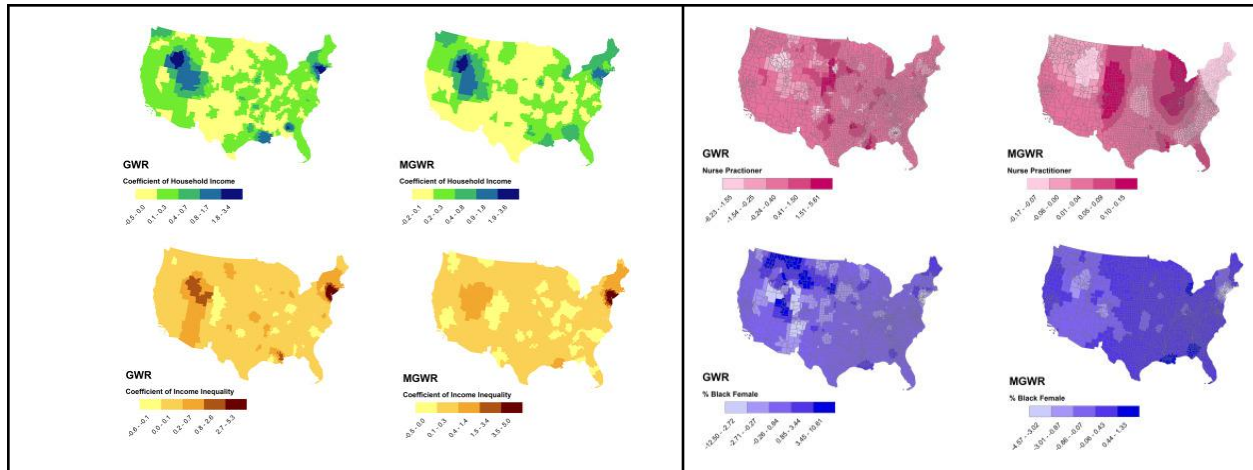
There are reasons why and how income affects any type of reported covid-19 measurements. Lower-income people live in denser areas, and in bigger households. During California's, in fact the country's, highest spread during the Winter/holiday season, cases surged exponentially because of gatherings. So I can see how larger numbers of people together for a decent amount of time, especially if living together, contributes to higher exposure and rates. Additionally, lower income neighborhoods typically have more essential workers, since many are blue collar workers who cannot do their jobs online like white collar workers in higher income neighborhoods can. Thus, more in-person interactions, even briefly, increases spread risk. Lower income neighborhoods also have lower-quality, and lower accessibility for good health care; death rates would therefore be higher in such areas since they are less likely to get proper care for cases where the patient really needs medical attention.

ARTICLES/ JOURNALS and INTERPRETATION

The first article, "*Disparities In Outcomes Among COVID-19 Patients In A Large Health Care System In California* " was published under the Health Affairs journal on 21 May 2020 and written by a number of authors who are listed in the introduction. They analyzed numbers at different hospitals in Northern CA, looking at suspected/confirmed cases and adjusted risk factors (age, sex, race, smokers, and income!) to create multiple predictor value regression models. They calculated and referenced "adjusted odds ratio" of each group's chances of hospitalization and death, in comparison to a reference value of 1. A 95% CI was used for each. For example, they would use median quartile income levels as a reference (obviously, ratio of 1) and find/compare the odds ratio of the other quartiles. Same goes for race (and etcetera), with white as the reference, and looking at calculated odds/ratio in other races. The results saw that some of the biggest differences were the 80+ age group (ratio of about 40x more likely to die) and income (in quartiles). The lowest quartile showed a ratio of about 2x higher and Quartile 3 had the lowest ratio of about 0.5, which is a four-fold difference (2 vs 0.5). The highest (4th) quartile (highest 25% of income) had a ratio slightly higher than that of Q3 (about 0.75x) but keep in mind that these are hospitalizations, not overall number cases or deaths. Hospitalization rates, as stated above, increase with older age, and Q4 (higher income) individuals generally are older in their career who earn more money.

The second article, "*GIS-based spatial modeling of COVID-19 incidence rate in the continental United States*," is written by Abolfazl Mollalo, Behzad Vahedi and Kiara Rivera under the Science of the Total Environment journal, published on 1 August 2020. They ran similar studies and processes as before. However, this would be a better study to align with my research topic, as it analyzes case rates, rather than hospitalization. My research topic entails rates of infection and spreadability in the population with income, and this study will map my goals more accurately. They analyzed dozens of factors within separate distinctive categories (demographic, behavioral, environment, topographic and socioeconomic) and used them as explanatory/prediction values for spatial, regression and log models. I will not be explaining their denotations or derivations for equations they used for each model, but I will be analyzing their summary statistics and GIS plots. Running their OLS (least squares) test, they sought to use income equality, median income, % of nurses, % of black females as variables to model incidence rates. Median household income rang up the highest coefficients compared with the other predictors, even more than income equality. This is the same result for the other models they ran and tested; and the biggest disparity for median household income (compared to other

factors it was grouped with) was shown in the spatial error model. Looking at the charts below, I can see that similarities when comparing multiple spatial models are most consistent for the green (household income) charts and also for the yellow (income inequality) charts, which doesn't fall far off.



CONCLUSION

The purpose of my study is to compare income (along with other demographic factors that income has influence on), in comparison to COVID-19 cases/deaths/positivity adjusted for population in county over state and state over the country. This is not to understand if income ranks as the best indicator, just if it is a relevant one to use. In this report, I used datasets including county, state and country level statistics for most COVID-19 measurements. However, I also used additional datasets from the USDA regarding household income in order to compare and attempt to give more insight on my report. A variety of questions (7 to be exact) asked helps answer and extract possible trends that these two factors could explain. 2 supportive studies were looked at and a comparative analysis/approach was done.

Generally speaking, it might seem as though income itself is not the very best predictor of my virus measurements. For example, in Questions 1 and 2, I do not see a trend comparing CA county case ratios during peak surges versus county-level income data. The same situation also arises in Question 4, which positivity rates in some high-income vs low-income states in the same geographical area. There was also not a clear trend for the state-level, as positive rates are a mere reflection of whether that area was an epicenter of outbreak during each period. However, the reason this is the case is that these comparisons use very large collections. In other words, I was using income from the entire county, which has large populations and thus, is subject to wide variations and unaccounted distributions. Mean household income would be more skewed compared to median values.

When looking at the state-level income data, there is even more inaccuracy versus looking at county-level data because an even larger scope with lower confidence/ representativity was used. However, when I analyzed neighborhoods within the city of Los Angeles during surging infection in the Winter, there is clearly a pattern between income and infection rates. The charts were extremely alike, in that high-income explains lower rates of infection while lower-income areas explain higher rates. Although income does prove itself to be a good indicator of virus measurements, it is best to use parameters as narrow as possible (such as neighborhoods instead of overall county, or a city instead of a whole state); this would produce more units (88 values for 88 cities in LA county vs. 1 value to describe the entire county) and to better account for disparities.

The articles assessed the relationship between income and infection statistics. From the first article, some important results were that age (80+ age group produced the highest odds ratio compared to the general population, ratio ~ 40x more) and income (the lowest quartile had a ratio 4x higher than Q3) were good variables to use for infection rates. Q4 (highest income; ratio of 0.75 compared to median) has a higher odds ratio for hospitalization/death compared to Q3 but that's simply because the highest earners tend to be of older age and more vulnerable. From the second study, a similar approach was used for case rates. The authors created different regression, GIS, etc models and considered the use of 35 explanatory variables in demographic, behavioral, environmental etc categories. Separate models, each using the same few variables and found that across the different models, determines that household income is one of the best predictors of infection rates or cases, even more than income inequality itself.

SOURCES

1. Azar, Kristen M. J., et al. "Disparities In Outcomes Among COVID-19 Patients In A Large Health Care System In California: Health Affairs Journal." *Health Affairs*, Vol 39 No 7, 1 May 2020, 17 March 2021, www.healthaffairs.org/doi/full/10.1377/hlthaff.2020.00598.
2. Mollalo, Abolfazl. Vahedi, Behzad. Rivera, Kiara M. "GIS-based spatial modeling of COVID-19 incidence rate in the continental United States,." *Science of the Total Environment*, Vol 728, 1 August 2020, 17 March 2021, <https://www.sciencedirect.com/science/article/pii/S0048969720324013>.
3. City News Services, N/A. "LA Neighborhood Map Details COVID-19 Cases and Deaths.", NBC Los Angeles. NBCUniversalMedia, 22 December 2021 [Accessed 18 March 2021].
<https://www.nbcboston.com/news/coronavirus/southern-california-coronavirus/los-angeles-la-neighborhood-coronavirus-cases-deaths-map-covid-19/2490987/>