Understanding County-Level COVID-19 Data

FLORIDA COUNTIES CASE STUDY

ANNA F HARVEY

Bellevue University

DSC 680: Applied Data Science

Prof. Catherine Williams

April 11, 2021

Abstract

Understanding the factors affecting transmission of COVID-19 has been a continually complex problem. This is particularly true in Florida, where thorough and clear data has been difficult to collect. Florida is a unique state with widely varying population demographics that are largely divided by county. Since the beginning of the pandemic, many of the counties have functioned under their own authorities, despite receiving different guidance from state and federal levels. This study investigated to what extent COVID-19 has affected Florida counties differently and if there is any consistent variance between counties with different population and population density.

Initial analysis of case percentage by population and deaths percentage by population shows that there is no strong correlation between population and cases or deaths. In fact, some of the lowest populated counties had the highest rates of cases and deaths. Further investigation into county demographics and COVID-19 mitigation policies is required to determine if any key variable affected these values.

Many key decisions have been made throughout this pandemic under the assumption that population is a main driving factor for COVID-19 transmission. This study suggests that this may not be the case. Further study into sociological variables, such as demographics, income, and behavioral trends, should be implemented to fully understand why different counties have different rates of transmission. This is vital to mitigating the effects of future pandemics should they arise.

Introduction

State-level data about the COVID-19 pandemic has been essential for state governments in their decisions about policy changes and safety restrictions. There has been little focus, however, on the differences of COVID-19 case and positivity rates between counties within the same state. States like Florida operate in a similar fashion to the nation as a whole, where there is an overarching authority that determines high-level policies, but the counties make many of their own decisions within state limitations. Data at a county level is crucial to analyze if local governments are meant to make decisions appropriate for their own communities.

The presumed driving factors behind the speed of COVID-19 spread and the severity of cases are population centers and age. Larger populations are assumed to have a higher rate of spread. Areas with older citizens are assumed to have a higher rate of death. Florida has 67 counties that range in population between approximately 8300 and 2.7 million[1]. There are several areas of the state with expansive retirement communities and elder care facilities. Additionally, as a political swing state, Florida is well-known to have a wide range of political beliefs. Since COVID-19 mitigation decisions have been largely influenced by the main political parties, pandemic behavior in different counties has also likely been influenced by the main party represented in each county.

Problem Statement

If population and age distribution play a significant factor in COVID-19 spread through communities, a strong correlation should be found between those variables. Should such a correlation exist, it would provide a more targeted strategy for different counties based on the demographics of the area. Alternatively, the lack of a strong correlation could suggest that other factors may be the driving forces behind spread and death rates.

Methods

Datasets were utilized from US Census data as well as the compiled county-level data collected by the New York Times and provided on GitHub[1,2]. The original county-level data includes daily reports from every county in the US from every day data was reported since early 2020[2]. Florida data was isolated from that dataset. The case and death numbers in the data are cumulative and each day's numbers are added to create a running total for each county over the course of the year. This data does not reflect the daily changes to case and death rates. The cumulative data for the most recent reported date was isolated to evaluate the overall case and death rates for each county. Features were created to standardize the case and death rates by population. The percentage was calculated by dividing the variables by the population. A standardized number per 10,000 people was also created by multiplying the percentage by 10,000.

Visualizations were created for the cumulative totals for the year to explore any possible correlations or outliers (Figure 1; Figure 2). It was determined that Lafayette County is an outlier for case rates with a significantly high percentage of 19% of the population infected since March 2020. Investigation into this situation revealed that there was a significant COVID-19 outbreak at the Mayo Correctional Institution in August 2020[3]. Miami-Dade County also could be considered an outlier for case rates at 16%. It is a surprisingly high number when compared to other large counties. Union County is the only outlier for death rates at approximately 46 of every 10,000 people dying from COVID-19.

Correlation matrices were calculated for all variables. Both Pearson and Spearman correlations were calculated, however, it is unlikely that any of the variables have a linear relationship therefore Spearman correlations should be primarily utilized. There are significant positive correlations between case percentage/death percentage (0.91) and standardized cases and deaths (0.91). Strong correlations between percentage and standardizations with their contributing variables (cases and deaths) were ignored. There are weak negative correlations between case percentage and population (-0.16). That was the strongest negative correlation between variables.

Results and Conclusions

There does not seem to be a strong correlation between population and the number of COVID-19 cases or deaths. If the number of people in an area were the primary factor in determining how quickly COVID-19 spreads in a community, we would see a higher percentage of cases in the highly populated counties. In fact, most of the counties with higher case rates are primarily lower population counties. We particularly see this in the smaller counties along the panhandle. Overall, we do see most counties clustered within the same range of case rates (between 6%-11%), regardless of population (Figure 2). The two exceptions to this are Miami-Dade County and Lafayette County. As previously mentioned, Lafayette County's primary factor in its high percentage of cases is due to the outbreak at the Mayo Correctional Institution. Interestingly, there are major correctional facilities located in almost every county, but not every correctional facility had a major outbreak. Miami-Dade County seems to display the assumption that a higher population leads to higher case rates. However, when comparing Miami-Dade to other large counties, the differences in case percentages are 5%-9%, while the other large counties only vary from each other by 1%-4%.

The relationship between death rates and population theoretically is that the higher the population, the more cases there are, and therefore the more deaths. However, since we see that population and case rates do not have a strong correlation, it makes sense that death rates and population do not have strong correlations either. Death rates and case rates have strong correlations to each other, which does confirm that theoretical relationship. The one outlier of Union County falls within the mid-level case rates for counties but has a difference of 14 more people per 10,000 dying from COVID-19 compared to the next highest death rate. Otherwise, the death rates for all counties stay between 6-32 per 10,000 people (Figure 2). That is a wide variance and further investigation into the variation between the main cluster is warranted.

Future Investigations

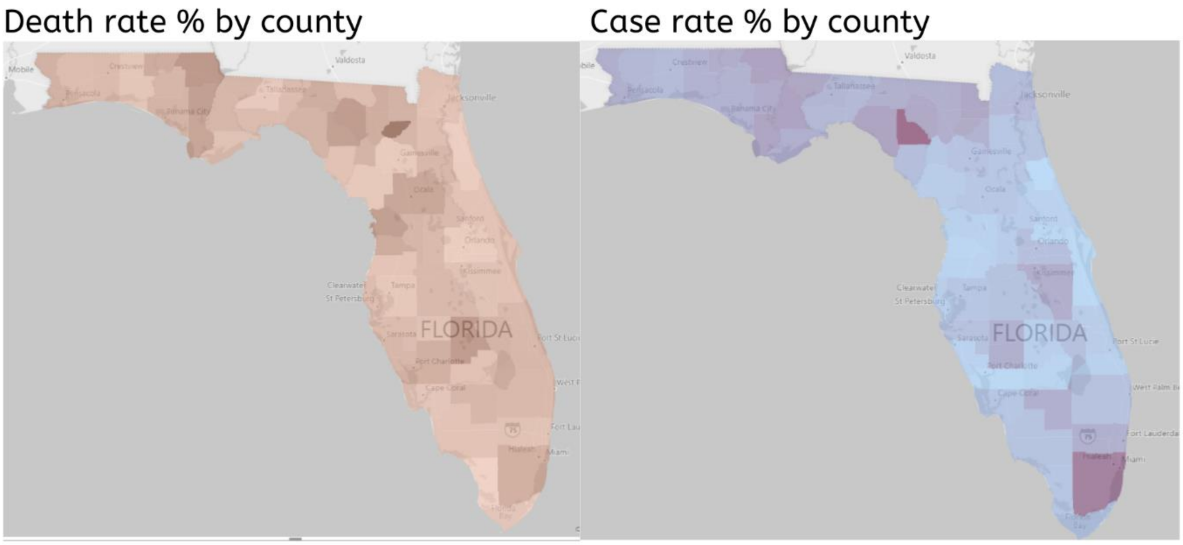
The affect of age distribution on case and death rates has yet to be investigated. The US Census data has some data missing from certain Florida counties and a more complete dataset needs to be found to accurately compare the counties' COVID-19 data to their age distribution. Miami-Dade

County’s high case rate does not have any variable strongly correlated to it within this dataset. Behavioral or other demographic information needs to be investigated to determine why the case rate there is so high. Union County also does not have a variable strongly correlated to its high death rate. Behavioral or demographic information need to be investigated for this county as well.

Other variables to investigate include the racial makeup of populations, income levels, access to healthcare, political diversity, behavior patterns, and dominant occupations in an area. It seems likely that the transmission of COVID-19 is not merely affected by the biological makeup of the virus. There are several sociological factors at play that could affect the way the disease spreads from person to person within a county and then to surrounding counties as well. It is important to investigate and evaluate these variables. Counties with high transmission rates will keep a state from maintaining control over a virus as people travel between counties. It also puts a strain on hospitals as people have to cross county lines to find open hospital beds. Ideally, significant sociological factors can be identified and addressed to prevent extensive disease spread in the future.

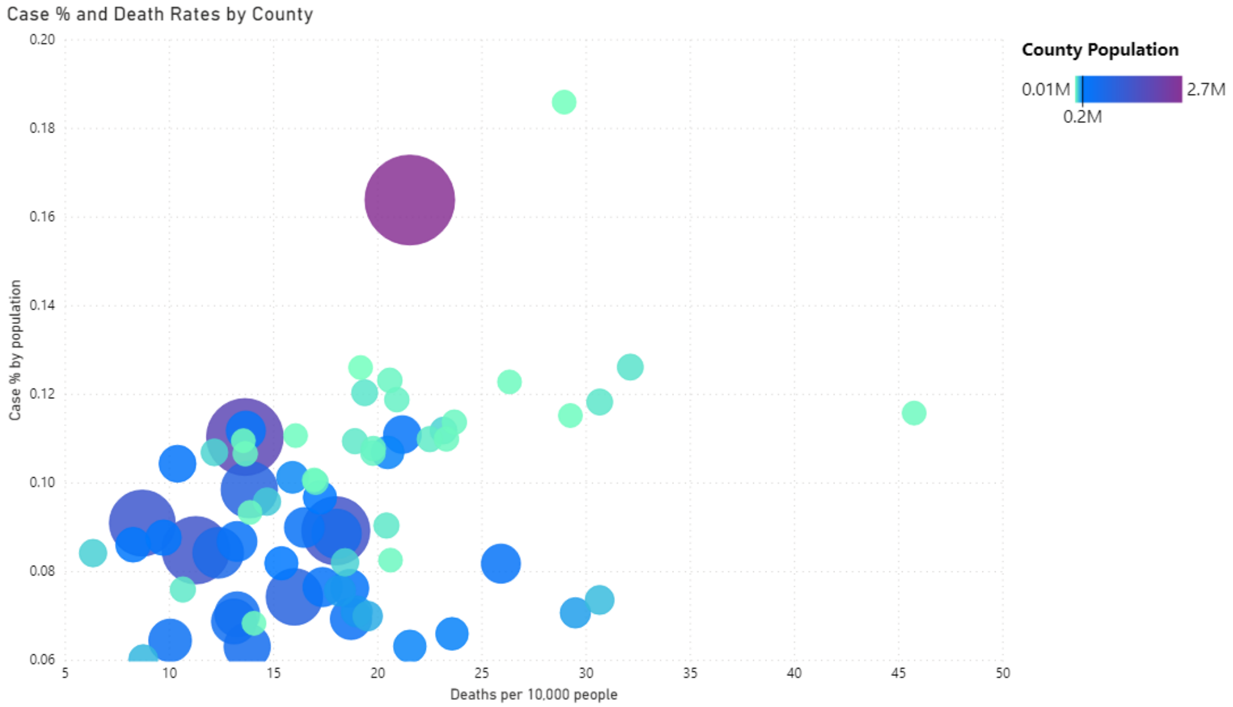
APPENDIX

Figure 1. Filled maps showing case and death rate percentages by county.


Death rate % by county

Case rate % by county

Figure 2: Scatter plot for case percentages and death rates with bubbles sized for population.


Case % and Death Rates by County

RESOURCES

1. *Florida Counties by Population*. Florida Outline. (n.d.). https://www.florida-demographics.com/counties_by_population .

2. Nytimes. (n.d.). nytimes/covid-19-data. https://github.com/nytimes/covid-19-data.

3. Renton, A., & Hayes, M. (2020, September 10). *Nearly 14% of the people in this Florida county have had Covid-19*. CNN. https://edition.cnn.com/world/live-news/coronavirus-pandemic-09-09-20-intl/h_1a0c7d2ff36b08ab4ca3ffbb1e304c8a .

https://github.com/anhar421/Portfolio

10 Questions

1. *Why did you use data compiled from Johns Hopkins University instead of data direct from the Florida Department of Health?*

   The issue of data transparency in Florida has been widely publicized. One of the main data scientists working for the state was fired for not manipulating data for the state's COVID-19 dashboard to promote reopening strategies. Florida has not been willing or able to provide detailed reports about COVID-19 cases and deaths in the state, much less at the county level. However, they do release daily counts for the counties and that data presumably is where the New York Times received their data for Florida counties. They collected the data daily over the course of the last year and have been able to compile it to show how the cumulative cases and deaths have changed over time. After spending a lot of time looking for appropriate data and coming up short in several official government venues, I decided to stick with the NYT dataset.

2. *What makes Florida different than other states when it comes to COVID-19?*

   Florida rapidly decided at a state level to abandon COVID-19 mitigation strategies. However, individual counties have vocally and continuously pushed against the state strategy and implemented their own mandates and policies. There have been attempts to thwart these efforts by the state, with varying success. What we essentially have is a state at war with itself and the citizen population is divided against each other, often along county lines. It is troubling but also provides interesting case studies to analyze how individual counties have been affected by COVID-19.

3. *What domain knowledge is required to have a complete understanding of COVID-19 spread?*

   Initially I believed that the main domain necessary for understanding COVID-19 transmission was the medical field. However, my analysis leads me to believe that sociological professionals should be consulted as well. Understanding the human behavior factor behind disease transmission is important and should not be dismissed if we wish to make better choices in the future.

4. *Which counties had surprising case and/or death rates? Why?*

   Miami-Dade County is the most surprising to me for case rates. It had a much higher spread relative to other high population counties. Someone pointed out to me that some of that may have to do with the large Cuban population centered in Miami-Dade County. There could be cultural aspects at play that affected the rate of transmission. What is also interesting is that although Miami-Dade had a significantly higher case rate than most other counties, the death rate was not nearly as high as many of the lower population counties with lower case rates.

5. *Why did you focus on population as a focal point?*

   One of the main arguments many have made throughout this pandemic is that higher population areas will have higher rates of transmission. This logically is true and we see that in the base numbers. However, before my analysis, I began to suspect that population may not be a consistent factor based on behavior and data I was seeing come out of Florida.

6. *What factors might contribute to population not being strongly correlated with case rates?*

   High population areas may have citizens who take more precautions when out in public. Low population areas may have less access to healthcare. High population density areas may also have low access to health care due to lower income populations. Lower income populations may not take the time to get tested when sick due to the need to be at work. The longer I have looked at this data, the more convoluted the sociological variables have become. There are potentially dozens of variables and interconnected variable combinations that could affect whether or not an area with a large population actually has high transmission rates.

7. *Why do you think county data has not come under more investigation when deciding policy?*

   It is difficult to standardize data across counties because in general counties come in a wide variety of sizes and populations. I think most of the government focus during this pandemic has been on trying to evaluate things in as broad a way as possible to try to keep things simple. Getting into the complexity of individual counties is something professionals really haven’t had time for during this emergency.

8. *Are COVID-19 case and death rates enough to make informed statements about the current state of the virus in a county?*

   I don’t believe so. I would also argue that it’s not enough to make informed statements about a state or nation either. This is especially true because each state reports its case and death rates differently. Without a global, consistent method of reporting, it is hard to establish thorough data about cases and deaths. Case rates are also dependent on people getting tested, which is a voluntary action. We have no idea what the actual rate of transmission is anywhere because no one can test every person in an area. Medical professionals have stated that excess deaths (the number of deaths higher than expected for the time of year) and hospitalization rates are better indicators of COVID-19 status. However, that does not directly take into account infections that are asymptomatic or mild. It is possible that those numbers could be inferred from the hospitalizations and excess deaths.

9. *Could COVID-19 case and/or death rates be predicted using predictive analytics?*

   I think it is possible, but it would be difficult. Sociological variables would need to be standardized in such a way to integrate them into machine learning or other predictive tools. It would also require some definitive answers on what sociological factors affect transmission.

10. *What future paths could you investigate to further understand COVID-19 from data standpoint?*

    Analysis of movement using mobile phone data would be a good step. Accurate assessments of county decision making on transmission rates would require a thorough understanding of each county's political policies regarding COVID-19. Studies into racial/cultural data and income distribution would provide more opportunities to see if those factors are significant.