The Effects of COVID-19 on a County, State, and Socioeconomic Level

Flora Dong, Eric Wang

Abstract

The outbreak of the novel coronavirus SARS-CoV-2, the cause of COVID-19, resulted in a global pandemic, ultimately leading to massive disruptions to everyday life for people of all backgrounds. This pandemic has inevitably resulted in numerous confirmed cases and deaths. International, national, state, and county level legislations are enacting their own restrictions in hope of containing the spread of the virus. In light of this situation, we analyzed the effects of county demography, socioeconomic status, and state orders to see whether these three different aspects play a role in the spread of COVID-19. Additionally, with the Linear Regression model, we explored whether the demographic population in an area affects COVID-19 transmissions or not, in an effort to capture the same conclusion as our analysis. In a series of three studies, we found somewhat significant correlation between urban areas and transmission rates, leading to state officials enforcing shelter in place (SIP) policies for residents to stay at home (SAH), ultimately slightly reducing COVID-19's incident, hospitalization, and mortality rate via social distancing.

Introduction

Coronaviruses are a group of viruses that cause respiratory tract infections that range from mild, the common cold, to deadly, like SARS-CoV-1 (SARS), MERS-CoV (MERS), and the newly evolving SARS-CoV-2 (COVID-19). In the early 2000s, SARS, caused by SARS-CoV-1 coronavirus, resulted in the 2002-2004 SARS outbreak, inflecting over 8,000 people and killing 774 people. Its successor, the new strain SARS-CoV-2 coronavirus, was discovered in late 2019, resulting in the coronavirus disease 2019 (COVID-19). The first known confirmed case of COVID-19 is believed to have occurred on November 17, 2019, infecting a 55-year old man with an unknown pneumonia-like infection in Wuhan, China. By December 31, the Wuhan Municipal Health Commission reported that they were treating pneumonia-like cases, in which a novel coronavirus was detected. Researchers believe that the coronavirus was first present in the Huanan Market in Wuhan as many of the first cases were linked to the market, making it possible that an animal source was necessary for transmission (source). Less than two weeks after the initial report, the first known death from COVID-19 was confirmed on January 11, 2020: a 61-year old man who had been a regular customer at this market. After efforts to try to contain COVID-19 within China, the first case outside of China was confirmed in Bangkok, Thailand on January 13, after a man travelled to Wuhan for a business trip (source). The virus soon spread to France, when the first known COVID-19 case in Europe was confirmed on January 24. The first case confirmed in the United States was a 35-year old man on January 20 in Snohomish County, Washington (source). COVID-19 continues to spread exponentially, resulting in the World Health Organization (WHO) declaring this the novel coronavirus a global pandemic on March 11 (source). Nearly two months have elapsed since this declaration, and COVID-19 continues to spread throughout 185 countries, infecting nearly four million people and killing over 270,000.

Questions We Explore

Given the limited dataset provided and its dated nature (the latest infection data is from 4/18/20), we decided to craft our questions around retroactively analyzing the effectiveness of various government policies and evaluating several common perceptions about COVID-19. We decided that the prediction of future COVID-19 spread is already well studied by experts. We drafted three main questions that we could explore that would provide great insight about the spread of COVID-19:

1. How does the population density and isolation of a rural town vs. an urban city impact their use of social distancing/SIP requirements? How has this impacted their infection rates? We know that population density plays a role in the spreading of a contagious disease as higher population density leads to more human interaction. However, it has been hotly debated as to whether or not small towns should be allowed to lift SIP orders while their urban counterparts remain on lockdown. Moreover, rural towns tend to have less access to the medical facilities that urban areas do. We aim to investigate if rural towns are truly better equipped to slow the spread of COVID-19 and if they are at an advantage for recovering.

2. How have state level approaches towards COVID-19 and their urgency towards the situation impacted their infection, hospitalization, and mortality rates? The federal government has since delegated the decision of social distancing down to the state level. It raises the question of whether or not an individual state's social distancing guidelines is able to influence the spread of COVID-19. Does an earlier SAH order mean less infection rates due to social distancing? Does SIP work in decreasing transmission?

3. Does socioeconomic status play a role in the ability to recover from COVID-19? If so, what role? COVID-19 has affected the entire world; however, it seems that some people are less worried about it. We have the Trump supporters rallying behind the reassurances of the President, largely

conservatives long wary of the big government feeling that accompanies forced quarantine. We also have billionaires like Elon Musk pushing to reopen his Fremont Tesla car manufacturing factory against public health recommendations and county orders. He defended his actions with a wide array of reasons including saying that he will be out on the production line risking his life with the rest of the workers (source). This raises the question: does socioeconomic status play a role in this pandemic? The United States has a privatized healthcare system that has long been accused of favoring the rich. We see reports of wealthy athletic associations receiving private tests inaccessible to the general public. Do the rich also have access to better medical treatment? Are they able to recover from the virus?

Description of Data

The Data 100 staff provided us with four complex datasets about COVID-19 as well as general statistics about counties that might contribute to the spread of COVID-19. The datasets are:

1. States: A general description of all US states as well as certain provinces/countries in the world. It contains general information about COVID-19 as it relates to that region, including, but not limited to, hospitality rate, number of confirmed cases, number of deaths, and mortality rate.

2. Counties: A collection of different socioeconomic statistics about counties around the world as well as other non COVID-19 health statistics. It also includes an ordinal date as to when certain states imposed different levels of social distancing restrictions.

3. Time Confirmed: A dataset containing the number of confirmed COVID-19 cases in each region at the given date.

4. Time Deaths: A dataset containing the number of deaths caused by COVID-19 in each region at the given date.

Additional datasets that were not provided we used include:

1. State Abbreviations (dataset): This is mostly a table that helps fill in some missing entries.
2. CDC SVI (dataset): Social Vulnerability Index. We use the Socioeconomic Theme as a guide for our wealth study.

Data Cleaning

For the purposes of our project, we will limit ourselves to the investigation of conditions within the US. The reasoning for this is two-fold. First, the US has a vastly different medical system and approach to COVID-19 compared to their counterparts with a national healthcare system, or those without a developed healthcare infrastructure, etc. Second, the data for most of the non-US locations are vastly incomplete. As such, adding them will force us to consider many factors and do a considerable cleanup of the data that is simply not worth our effort. Evidence of the lack of sufficient data can be seen in the table below.

SVI dataframe: We extracted SVIPercentile summary statistics using the clean.py script provided by the Yu group (script).

States dataframe:

- We dropped the following columns:
    - 'Last_Update': Every entry was the same as we only had one update
    - 'Lat', 'Long_',: We are not considering the locations as we are ill equipped to model pandemic outbreak networks. Moreover, no county can simply decide they want to move to a different location. Instead we will be focusing upon the efficacy of policies the government can control such as social distancing.
- We filtered the data to only contain US states + DC
- We removed the Active Cases column along with recovered column due to incomplete data
    - 22/51 states did not report recovered cases
    - Indiana, Nebraska, and Nevada are noted to not have hospitalization rates
    - North Dakota, South Dakota, Iowa, Nebraska, and Arkansas are noted to not have SIP dates

Counties dataframe:

- We dropped empty rows with the following identifying factors:
    - Null StateFP, countyFIPS of City1 and City2, Shannon SD
- Bedford, VA is kept but has a lot of missing data
- All ordinal time entries are copied as Timestamps into another column for easier data visualization
- One Hot Encoding: CensusRegionName, CensusDivisionName, Rural-UrbanContinuumCode2013 (these definitions can be found here)

Timeseries dataframes:

- We dropped the following columns:
    - UID, iso2, iso3, code3, FIPS, Admin2, Lat, Long_
- We extract a county column from the combined key.

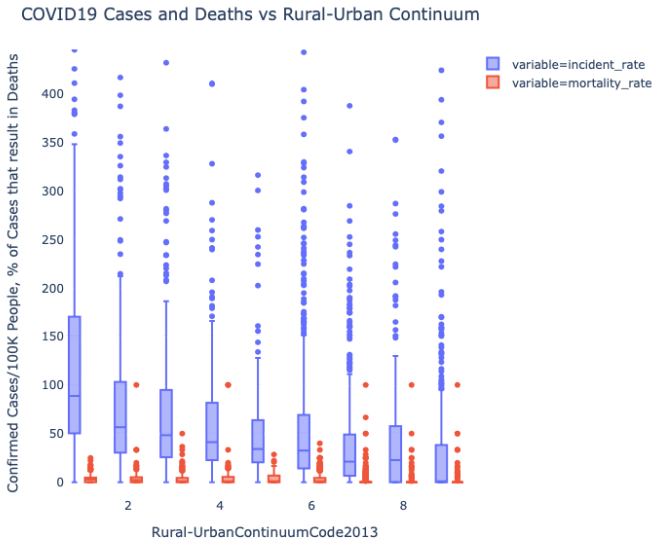
Full_stats dataframe: A meta dataframe that combines data represented at the county level (all the dataframes except states are included)

- We note that we used an "inner" join on FIPS codes to eliminate as much missing data as possible
- We also note that there are around 100 missing entries as a result, mostly due to "Unassigned" and "Out of State" entries. Please refer to the notebook for a detailed breakdown of entries that were dropped.
- A "limited" dataframe can be generated as well that only includes the bare minimum timeseries data required for feature engineering

Feature Engineering: We engineered several features:

- Incident rate of cases and deaths relative to the population: This is motivated by the reasoning that larger cities with more people will naturally have more cases as there are more people to infect. By scaling the number of confirmed cases and deaths by the population of the county, we are better able to compare counties with different populations.

A)



B)

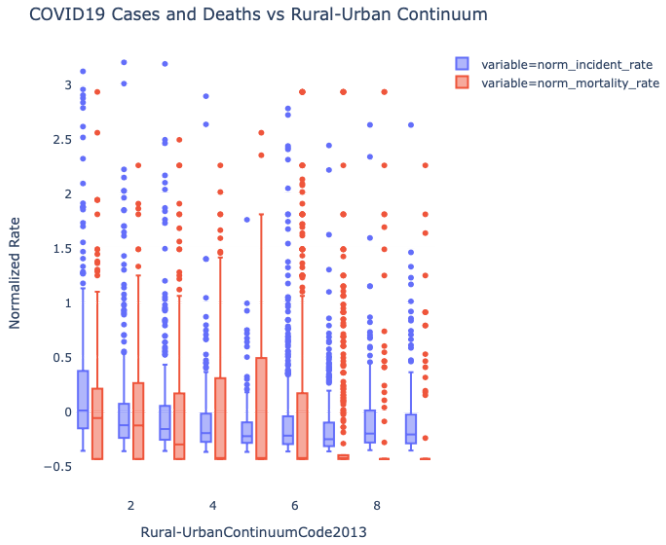


Figure 1: A) Unnormalized incident rate and mortality rate from COVID-19. There is a greater incident rate for urban counties, but unnormalized data is hard to interpret. B) Normalized incident rate and mortality rate from COVID-19. This shows an interesting mortality rate interquartile range (IQR) that peaks around the suburban areas, an observation not present in 1A.

- Mortality rate: A person not infected by COVID-19 cannot die of COVID-19, as such, we scale death counts by the number of cases in the county to better evaluate the ability of a person to recover.
- Growth rate of confirmed cases over a 14 day interval: There are existing models suggesting that diseases like COVID-19 spread at an exponential rate. What this suggests is that instead of calculating the number of cases that have occured, we should really be asking how fast the cases are spreading. The growth rate provides a transmission rate that we can use to better evaluate the effectiveness of various enforced policies.

Analysis

Urban-Rural Study

When considering states and their average Rural-Urban Continuum code, we see a correlation between incident rate and their continuum code. We also see a weak correlation for mortality rate. This will act as guides for the rest of our study at a county level. We also plot hospitalization and testing rates but find that they aren't very correlated at a state level. We will investigate other factors that may influence these rates in the States specific study.

Looking at the county level, we attempt to plot incident and mortality rates as a box and whisker plot (Figure 1A, B). We quickly see that the Figure 1A is incredibly hard to interpret. Clearly we can see the incident rate decreasing as we approach the rural counties. However, our y-axis scaling did not fit both incident rate (ranging between 0 and 100,000) and mortality rate (ranging between 0 and 100), making it hard to interpret mortality rate without zooming in further. As such, we decide to plot the normalized data from here on out. In Figure 1B, incident rates decrease from urban counties to rural counties. However, in terms of mortality rate, it is far less clear. The medians still seem correlated to the continuum code, but the 3rd quartile and upper fence actually seem to peak with a continuum code of 5 (urban population of 20,000 or more, not adjacent to a metro area).

Furthermore, we try to see if we can identify trends in confirmed cases growth rates, our other engineered feature. As seen in Figure 2, 14-day growth proportions (norm_growth_14_confirmed) are surprisingly low for urban counties, rather peaking in rural counties. This suggests that a boom in COVID-19 cases between 4/4 and 4/18 occurred in rural counties. Looking at the 31-day growth rates (norm_growth_31_confirmed) and the 14-day growth rate as of 4/1 (norm_growth_14_31_confirmed), we see the expected pattern of urban areas having a higher growth rate than rural areas. This may be caused by the virus taking longer to reach rural counties as urban areas are often ports of entry for patient zero, such as a businessman traveling to Asia. In comparison rural counties only get impacted when an infected traveler from the city passes by. As such urban areas are already reaching peak infection rates and imposing strict social distancing standards, rural areas are still approaching that peak or imposing those standards (source).

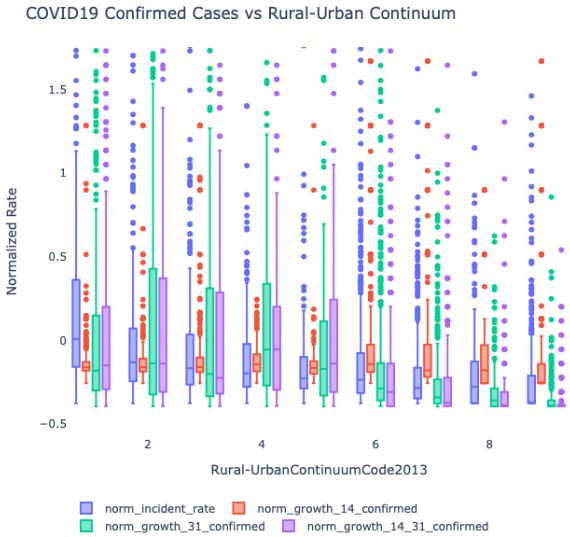


Figure 2: 14-day growth proportions between urban and rural counties. Growth proportions differ based upon sampling time, resulting in low proportions for urban counties and peeking in rural counties, suggesting that a significant increase in COVID-19 cases occurred sometime between 4/4/2020 and 4/18/2020. Looking at the 31-day growth rates and the 14-day growth rate as of 4/1, we see the expected pattern of urban areas having a higher growth rate than rural areas.

Taking a look at social distancing start dates vs. Rural-Urban Continuum Codes reveals the expected trend where cities tend to impose restrictions first before the rural counties followed suit. It should be noted that in Figure 3A, B, the violin plots have such a large overlap that nothing conclusive can be deduced. The medians, however, do follow said trend, but the quartiles and fences are well within one another, so these features are not that dependent on each other. This potentially suggests that the growth of confirmed cases over the 4/4/20 to 4/18/20 was caused by something other than a later social distancing start.

Let us take a closer look at this by plotting social distancing directly against the growth rates. In Figure 5, we see that as suggested previously, not much can be said conclusively. Even when looking at SAH orders (Figure 3D), we see two different peaks in the IQR, suggesting that there isn't a clear relationship stating that earlier SAH measures are more effective. Perhaps one can argue that there appears to be more consistently large IQRs for dates after 3/29; however, we would need more data to say for sure. Looking at >500 gatherings (Figure 3C), we see the same sort of inconclusiveness with 3/23 being an abnormal peak, but all the dates after it seem to have comparable or better IQRs compared to those before 3/23.

States and Rates Study

Dr. Anthony Fauci, the director of the National Institute of Allergy and Infectious Diseases, started using the phrase "flatten the curve" when healthcare systems were being overwhelmed by number of COVID-19 cases exponentially rising, referring to the curve that represents the number of cases that occurs at once. In attempts to flatten the curve, public health officials recommended everyone to practice "social distancing," a term used to describe the act of maintaining at least six feet away from the nearest person not in the same household. Individual states officials started to implement SAH orders, requiring all residents to SIP, essentially quarantining at their primary residence. We want to analyze when the SIP order was initialized in each state and its effect on the incident, mortality, and hospitalization rate.

Looking at just data within the US, we extracted only the necessary data from the counties dataframe. To make use of the SIP times, it was necessary to convert ordinal time to readable dates. Upon merging with the states dataframe to gather the different rates and grouping states together, we finally arrive at a dataframe that gives us each state's SIP date. Make note that there are five states that never issued a SIP: North Dakota, South Dakota, Iowa, Nebraska, and Arkansas.

Upon initial elementary analysis, we notice that New York was a significant outlier as it is affected exponentially worse than the rest of the country. In an attempt to generalize this data to the majority of the US, NY was dropped and we proceeded with the remaining 44 states plus DC.

Because we wanted to see the SIP date's effectiveness, it was necessary to sort the states according to their SIP date. The greatest problem we had with this was rather what we would do with these dates and rates. There were so many different factors that played into a state's infection rate, and some dates had significantly more states than others, providing an unbalanced dataset. It seemed as if these rates each played their own role. However, we took both the mean and median of each rate at that date – we wanted to see which value

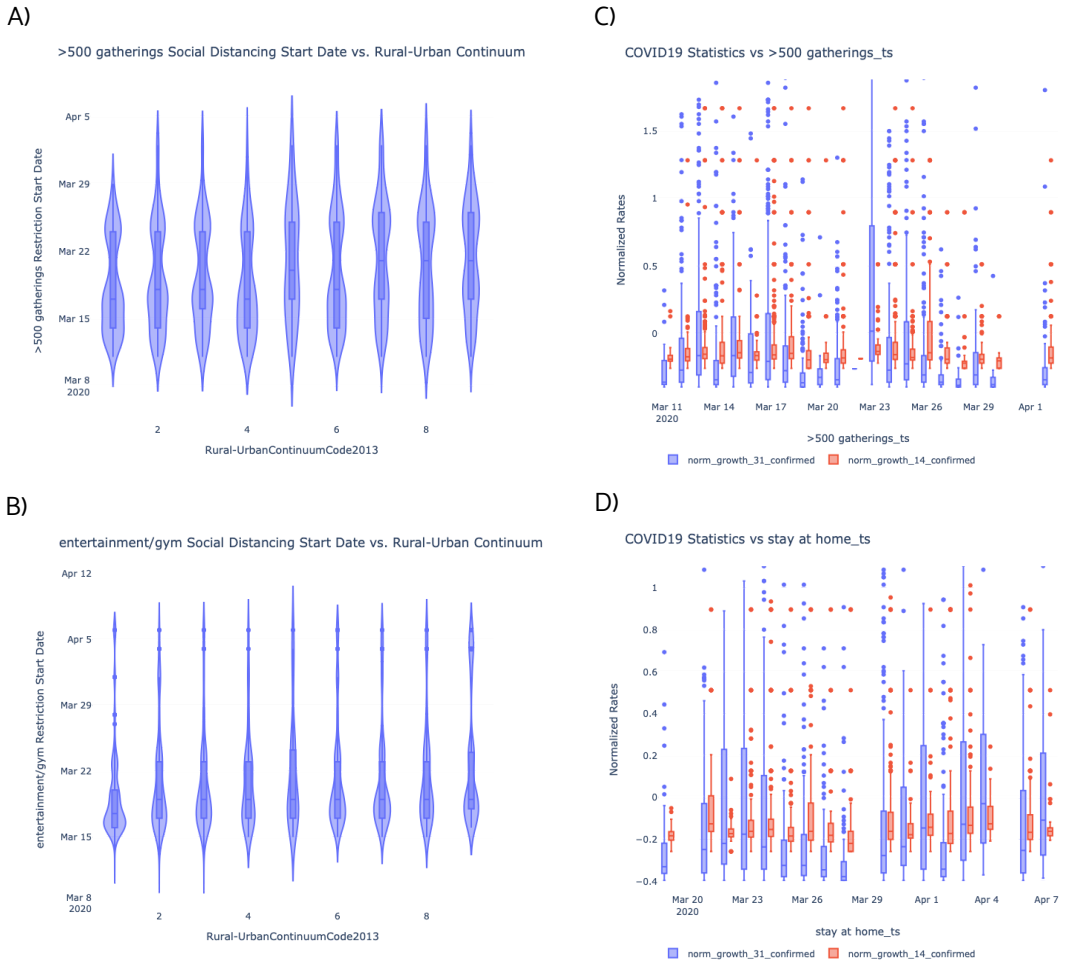


Figure 3: A) Shows social gathering restriction start dates on events with more than 500 people, suggesting a slight trend for earlier start dates in urban counties. B) Restrictions on entertainment and gym venues also started earlier in urban counties. C) Normalized 14 and 31-day growth rates don't appear to be affected by the start date of restrictions on large gatherings. D) Likewise SAH restriction start dates don't make a clear impact on 14 nor 31-day growth rates.

would provide us with a greater correlation. These two values were used for the respective SIP analysis.

Our first step was to see if there was a correlation in each of the individual rates. To do so, we plotted the corresponding median and mean of each rate at each date on a bar graph to visually see if there was a trend (Figure 4A, B). Upon doing so, we noticed that there was not a significant enough difference between mean and median values, thus resulting in us proceeding solely with the median values, as this is more representative of the true middle (Figure 4C). At first glance, each rate seemed to have their own correlation: the mortality and incident rate seemed to have a slight negative correlation with the SIP date, but hospitalizations seem to stay constant, prompting us to compare all three rates together.

It can be reasonably assumed the incidents, hospitalizations, and mortalities are positively correlated between one another: the higher the incident rate, the higher the hospitalization rate should be – more people are getting sick, more severe cases; the higher the hospitalization rate, the higher the mortality rate – more number of severe cases, more likely people these cases turn fatal. Normalization was an effective approach in comparing these three values since their rates per 100,000 people were quite different. Upon normalizing each rate, we analyzed how many standard units (SU) each date's rate was off from the median once again in a bar plot (Figure 5A).

Figure 5A provides us with a better visualization for comparison. In terms of mortality rate and incident rate, we see that their respective values in SU decrease as we approach April – these rates decreased with the later SIP dates. However, interestingly enough, although the hospitalization rate exhibited a similar behavior, it was noticeably more level, with equal number of values above and below the mean.

Figure 4: A) Average incident, hospitalization, and mortality rate by when SAH orders were enforced. No clear trend between the three rates. B) Median incident, hospitalization, and mortality rate by when SAH orders were enforced. Again, no clear trend between the three rates. C) Overlay of average and median incident, hospitalization, and mortality rate by when SAH orders were enforced. No noticable difference between average and median.

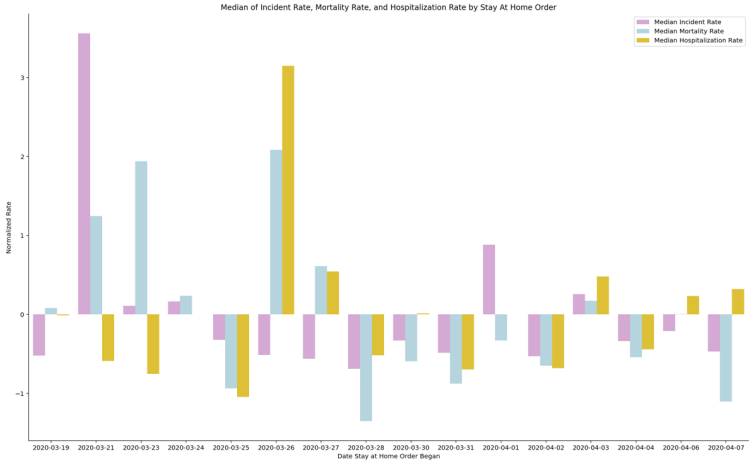
We want to point out the significant spikes on 3/21 and 3.26. We see that Illinois, New York, and New Jersey implemented an SIP order on 3/21 while Kentucky implemented an SIP order on 3/21. Despite removing New York from our data, we see that New Jersey, the second hardest hit state, as well as Illinois, the state with the third most cases (as of 5/13), still skewed our data. On the other hand, the increase from Kentucky is quite interesting, as no data suggested its rates to be out of the ordinary. However, this is just an initial glance from the naked eye; a more in-depth analysis would be required to further hone in on this idea.

As such, we decided to plot the normalized data on a scatter plot, giving us the ability to draw the best fit line (Figure 5B). This scatter plot reinforces our observations from the aforementioned normalized bar plot. We can distinctly see a negative-sloped line for both the incident rate and mortality rate (albeit the fact that mortality rate's slope is more negative than incident rate's slope). However, the hospitalization rate's slope is near zero, if not slightly positive. This is an unexpected observation as previously, we hypothesized that these would all be positively correlated, all exhibiting a positive slope or a negative slope. We considered calculating the slope of each respective regression line to further reinforce this idea, but such calculations would require significant work for little additional information.

The negative slopes in incident rate suggests that the later the SIP date was enforced, the fewer cases there were. Although this was not one of our initial hypotheses, this led us to wonder why states enforce a SIP mandate in the first place. In America, it is evident that state officials only act out of absolute necessity – if a state does not have many COVID-19 cases, COVID-19 does not affect their communities as much, and thus, state officials will not issue such SIP order. The later the SIP date is in effect suggests that the state was not particularly hard hit at the start of the pandemic, thus resulting in incident rates that are lower than states who had to issue such SIP order early on due to their growing number of cases.

With these initial results, it seems that state regulations had a very slight effect on the incident and mortality rate. This can be seen with the negatively sloped incident and mortality rates as time went on.

A)



B)

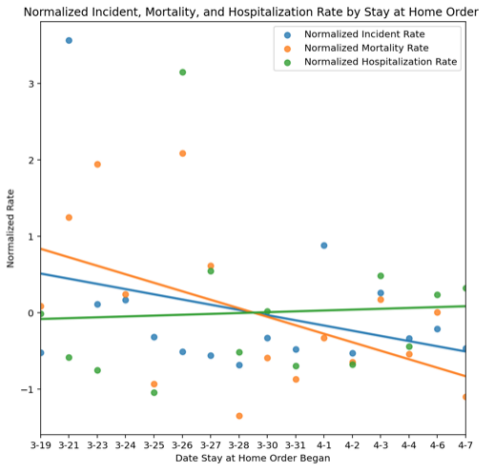


Figure 5: A) Normalized incident, mortality, and hospitalization rates amongst enforced SAH all dates. There does not seem to be much correlation between the rates, but perhaps a slight decrease in the normalized value as time goes one. B) Another visualization of 5A to see potential correlation via line of best fit. Clearly, there seems to be a negative slope in the incident rate and mortality rate, but no correlation/slight positive slope in hospitalization rate, suggesting an interesting unexpected inverse relationship between incident/mortality and hospitalization.

However, on the contrary, the hospitalization rate was not affected in a positive manner, perhaps even negatively affected, suggesting that other factors, like genetic mutation of the virus itself as well as socioeconomic status and accessibility to healthcare, have a greater role in hospitalization rate than merely when SIP was enacted. With this said, it is interesting to see that despite early efforts to minimize the spread of the virus and flatten the curve, this SIP method of encouraging the general public to practice social distancing had little effect on the actual infection, hospitalization, and mortality rates – the earlier the SIP was enforced, the higher the rates generally were, perhaps suggesting the idea that SIP is not as effective as public health officials have stressed.

Overall, it seems as if although state regulations have the slightest effect on the spread of COVID-19, it is generally inconclusive as to whether the SIP order had an overall large effect. We ultimately need to consider different features that likewise affect these rates.

Socioeconomic Study

Plotting the normalized mortality rate against overall Social Vulnerability Index (SVI) ratings and SVI's Socioeconomic (SE) theme reveals that there is no correlation for mortality rate.

However, our modeling step shows that both of those are important for predicting norm_growth_14_confirmed. As such, we have plotted that as well. Unfortunately, there doesn't seem to be any correlation other than a few outliers on the higher end of the SVI's SE theme spectrum, suggesting that further data cleaning, handling, and investigation is required. Most likely this is due to the calculation of growth_14_confirmed.

Modeling

For our model, we attempted to validate the features that were explored above. In particular, we explore the Rural-Urban Continuum Codes, Population Density, Social Distancing Guideline Start Dates, and SVI ratings and themes. To begin, we looked to model several statistics: incident rate, mortality rate, and 14-day growth rate. We chose to evaluate our model with a Mean Absolute Error scoring scheme. Our reasoning for this is that it doesn't get adversely affected by outliers like New York; however, it is flexible enough to accommodate the large number of counties with 0 cases. If we had chosen to take the median absolute error, it may very well report 0 as the median error even if all we did was predict a constant 0, depending on how the train test splits go.

We created a baseline model that predicted those aforementioned statistics based upon a linear regression of the 3/18 time series data. This serves as a good but naïve estimate for incident rate and 14-day growth rate (mean error ~0.36 SU). However it performs pretty poorly when it comes to mortality rate (mean error ~2 SU). Future models are scored against the baseline model by finding the difference in cross validation scores.

Next, we try to see if the features we studied can outperform the baseline model we provided. We called this the target model. We broke this down into multiple steps. First, we created one large model that took into account all the features. This model performed just a tad better than the baseline model with a mean error of -0.35, a 0.01 SU improvement. We decided to study each feature set on its own, breaking the features along the lines of Urban, Social Distancing, and SVI. Running these models against the

baseline we got improvements of 0.01, -0.01, and 0.00 SU respectively. So maybe they weren't actually improvements but this seems to line up with what we had explored above (Urban being a slight predictor and Social Distancing and SVI generally didn't help).

This is rather unsatisfactory. It would be nice if we can have a predictor that can beat 1-month old time series data. As a result, we attempt to use the SelectKBest filter and place the k features exhibiting the most influence. We removed restrictions on the features available to the model as long as it is not time series data or state names and coordinate data (reasoning outlined in the data cleaning section). We also decided to focus only on the 14-day growth response variable. Running the baseline and our target models we got errors of -0.3 SU, once again in line with our prior experience that our target model really didn't make a difference from our baseline. Running the SelectKBest filter we get that the top 10 "best" features involve pre-existing conditions, smoking status, number of hospitals, SVI data, and Urban data. This is interesting as pre-existing conditions are said to have effects on mortality rate but perhaps they also affect one's susceptibility to COVID-19 or one's response to it would be severe enough to be given one of the limited test kits.

We then tried to tune the number of features to select by conducting a grid search supported by cross validation. This landed us with 2 features: 'RespMortalityRate2014' and 'SVIPercentileSEtheme'. This seems to make sense as COVID-19 is a respiratory disease and early reports signaled that smokers and vapers were particularly susceptible to it. What is surprising is that SVI SE data being used seems to go against the findings we had in the Socioeconomic study. Moreover, it is surprising that the model is best fitted with just two features. This further highlights how the data is probably not best suited to answer the questions we have proposed or perhaps linear regression would benefit from regularization in order to prevent overfitting with more features. This is something that can be further investigated as an extension of this research. This model scored a CV score of -0.28 which is 0.02 SU better than our other models.

Overall, the models that we have created performed roughly the same. These models served to reinforce the results of our data analysis that suggested little to no correlation between the explored features and COVID-19 statistics. Assuming a Gaussian distribution, on average we were within 75% of the data – not terrible but also not great. In the future, we should look at more features outside of this data set along with different regression models that can better model a pandemic. It would be interesting to model the pandemic as a network of individuals with various transmission rates depending on social distancing guidelines, population density, etc.

Reflection

(i) What were two or three of the most interesting features you came across for your particular question?

The most interesting and unexpected feature we came across was growth_14_confirmed, which showed that rural areas are currently under a confirmed case growth spurt whereas urban areas have a slower growth rate. We predicted that urban areas would always have a faster growth rate due to their density, but it was interesting to see the inverse relationship between rural and urban growth rates.

Additionally, an unexpected feature we found was the relationship between incident, hospitalization, and mortality rates. It seems logical these three rates would be all increasing or decreasing as one leads to the other, but what was most unexpected was the "middle" one in this cascade, hospitalization, was the one that was inversely related to the other two. It seems that the mortality rate decreased as people got more severely sick, but perhaps these hospitalizalized patients all miraculously recovered at a greater rate than before. Whatever the reason may be, it was certainly unexpected.

(ii) Describe one feature you thought would be useful, but turned out to be ineffective.

Wealth doesn't seem to actually make an impact, which is surprising given the privatized healthcare system we have. We hear news of wealthier Americans having more access to test kits and such (NBA basketball players all being tested for example). But at the same time our data is relatively limited. While "rich" counties don't seem to be doing any better than their poorer counterparts, we cannot say that "rich" individuals are not doing any better. However, this could be slightly biased as poorer counties may not be fully reporting their COVID-19 deaths; due to their lack of accessibility to COVID-19 tests, there may be deaths caused by COVID-19 where the victim doesn't even know that they have contracted this virus in the first place. Our indicator for wealth was the CDC's SVI dataset, on particular the Socioeconomic Theme of that data set.

(iii) What challenges did you find with your data? Where did you get stuck?

Having these large datasets, we were unsure of what direction to go in. The first few ideas we had resulted in dead ends that were completely inconclusive. Our greatest challenge was tackling large scale questions that were answerable given our datasets but also encompassed enough breadth that allowed us to go in a multitude of directions. We got stuck here at the very start as we did not know what direction we wanted to proceed in. They say getting started is the hardest part of any project, and this certainly was the case for us. Data cleaning was another challenge, as it was extremely annoying dealing

with the numerous ways of documenting the same thing. Finding a consistent format for our data was challenging nevertheless; however, once we overcame such a hurdle, EDA and in depth analysis was much smoother.

(iv) What are some limitations of the analysis that you did? What assumptions did you make that could prove to be incorrect?

We had to limit most of our analysis to just the US as most of the data outside of the US was sparse and contained a lot of null values. We could have attempted to manipulate this null data accordingly, but such manipulation would potentially introduce biases, noise, and other undesired artifacts, so we decided to continue with just the US data as we already had most of the values. It would have been interesting to see how the US compares with other countries that have a similar economic status and were also greatly affected, like China, Italy, France, and Spain.

The US has been notoriously known for being behind on testing, so despite the fact that we assumed the data to be an accurate representation of how many cases/deaths there are, we know that this is an underreported value and that in actuality, the US has thousands of cases not being reported, resulting in an underreported number of deaths as well. Additionally, we assumed that the minimum number of confirmed cases was at least 0.1 and not 0, since dividing by 0 would give us an undefined value. However, this means that when a region that originally had no cases confirms a case, their growth was by 10 (0.1 to 1) instead of infinity (0 to 1). This suggestion was probably off from the true growth factor as well. Lastly, we only used data until April 18, and with situations changing literally every day, the data since April 18 definitely has new trends and results that could completely comply with our findings or go against what we have found. There has been numerous data and findings since April 18 that we did not consider, and in order to understand the complete picture, we need to consider the most up to date data.

(v) What ethical dilemmas did you face with this data?

Analyzing this data as the global pandemic is occurring is kind of a two-fold situation. On one hand, it is necessary that we understand the patterns of this virus so that we can accurately model and predict future scenarios and limit the number of cases and deaths in the months to come. However, many people are suffering and losing their lives because of this virus, and this is something we do not and cannot take lightheartedly. We are fortunate enough to be healthy to conduct such an analysis, but seeing these numbers sheds light on the numerous families that have lost a loved one due to this virus. Moreover, we do not evaluate the economic impact of the social distancing orders. As such, suggestions that social distancing works should still be weighed carefully against the economic loss and its implications for people's livelihood.

As mentioned above, this data given (and COVID-19 statistics about the US) is not representative of the true value of US' COVID-19 numbers as they have not been adequate testing. As a result, there is bias present in the data, presumably biased towards the wealthier Americans that have stable healthcare insurance and thus, access to necessary materials. Because the data has these underlying biases, the inference drawn from such data are also based on said biases and may exhibit similar biases.

(vi) What additional data, if available, would strengthen your analysis, or allow you to test some other hypotheses?

Most of our analysis was done with solely US data. For the US, if we had more individualized data about each patient, we could have analyzed more about mortality/hospitalization rates, especially the role that pre-existing health conditions play.

We felt that socioeconomic status should have played a more significant role than we concluded; most situations like this reveal the privileges the affluent class has, namely the fact that the privatized US healthcare system is more tailored to these people – they potentially have the luxury to purchase test kits and private hospital rooms. Having more socioeconomic data would have allowed us to highlight these disparities between different socioeconomic classes.

Governor Gavin Newsom was the first to enforce a SIP order for California, effectively doing so on March 19 (source). Our data only goes up until April 18. This is less than a month's worth of data, perhaps even less for states with later SIP dates. If the US citizens actually comply with SIP orders, perhaps a few more months of quarantine may yield positive results that stem from such social distancing. On a related note, despite statewide policies, some counties simply don't care and do not put an effort in enforcing these SIP politics, ultimately compromising the accuracy of our data. Having data that shows whether or not a county is truly following social distancing guidelines may provide us more accurate results of infection, hospitalization, and mortality rates.

Having more international data would have allowed us to extend our analysis further to other countries, providing a foundation for us to compare and contrast US's statistics with other countries. This will allow us to answer questions regarding the US government's response along with the efficacy of the US's unique healthcare system.

(vii) What ethical concerns might you encounter in studying this problem? How might you address those concerns?

The majority of the data here comes from the decennial US Census. As perfect as it tries to be, a significant number of people residing in the US are not counted for, like certain Native American tribes, undocumented immigrants, and the homeless population. Unfortunately, these are also the groups

that are most susceptible to contracting the virus. We could potentially gather more accurate data to combat these concerns, but federal policies make it difficult to accurately analyze such populations.

The US is a very diverse mixing pot of people of all cultures, backgrounds, and socioeconomic status. During our analysis, we generalize to a great extent, eliminating a lot of the diversity, which is not only not representative of the full picture but also unfair to the US' diverse population. Doing analysis on a more precise and specific level is possible, but this would require a lot more time, energy, knowledge, and data resources.

Like mentioned above, the pandemic is current and ongoing – we must be very sensitive about the fact that we are doing analysis on a very grimm situation that the world is currently facing. We all understand the gravity the current world faces and the fortunate position we are in to be able to study this data.

Future Directions

Through our COVID-19 analysis, we analyzed the impact of rural vs. urban areas when it comes to COVID-19 cases. While there seems to be a trend in the box and whisker plots in both incident rate and growth rates of varying durations, this did not play out in the model, with the SelectKBest selector opting to select other features instead. Moreover, the model with the urban features did not perform significantly better than the baseline. Future analysis as to why this occurred may lead to further insights on best pandemic modeling practices along with a more detailed explanation as to why the 14-day growth and the 31-day growth rates were so different.

Our current analysis on states' SIP dates and their respective incident, hospitality, and mortality rates is very generalized as we use the median values for all dates. Perhaps incorporating other social distancing factors like compliance and population density may provide us a better picture as to whether the SIP ordinance is actually effective or not. Our analysis is on the state level; however, we are given information on the county level. We could also zoom in on certain counties and the SIP effects on each county. Did counties with a higher population density show a greater improvement in COVID-19 cases after the SIP order was enforced or did population density not play a major factor in COVID-19 transmissions? In fact, the first counties (Alameda, San Francisco, Santa Clara, Marin, Contra Costa, San Mateo) that were ordered to SIP occurred even before the first official state did so. Did county level orders play a larger role than state level ones? Or did the state officials have a greater say when it came to SIP?

We all heard about all the NBA basketball players being tested within days of the first NBA player testing positive for COVID-19 despite there being a huge national shortage of testing supplies. How did these players miraculously get their hands on these tests when the general public had been trying to get tested for weeks? Additionally, one of the first celebrities to confirm to have tested positive for COVID-19 was Tom Hanks and his wife Rita Wilson. Shortly after that announcement, a flood of celebrities started to announce that they had tested positive for COVID-19 – Boris Johnson, Sophie Trudeau, Chris Cuomo, Daniel Dae Kim, Idris Elba just to name a few. Although our data didn't show much difference between socioeconomic classes, the foundation of the American healthcare system is completely biased towards those that are affluent; it is inevitable that during a health crisis like this that such biases will shine through. How does having a larger income play a role when it comes to getting adequate equipment and tests? How does fame and celebrity status affect this?

Moreover, the model can also be further improved. Currently, we are using a linear regression model; however, pandemics definitely are not that simple. As such, future research should look into more realistic models, such as the infection network model where we simulate the coronavirus infecting the people a patient interacts with. By changing the social distancing, population density, and socioeconomic factors, we can simulate how the infection rate may change. Furthermore, in this paper, we largely disregard the time series data, opting to use only select dates for our analysis. We can further improve the accuracy of our models and analysis by looking from a day to day perspective and utilizing the full time series dataset. This can take the form of average day to day infection growth rates to modeling how the rates change over time with respect to social distancing orders to using each day as a "correction factor" to help tune a simulation model that simulates each day of the pandemic. There are a lot of future directions that we can take with this data, but the analysis we did was certainly rewarding! We would love to see a compilation of all the directions proposed by Data 100 reports.

Thanks to the Data 100 Staff for an amazing semester! We appreciate all your efforts in managing this course during a pandemic. Hope you all stay healthy and happy :)