# Industrial Makeup of US Counties: Impact over Time on Spread of COVID-19

Athan Liu, Joel Whittier, Ishaan Bhojwani

June 12, 2020

# Contents

# 1 Introduction

## 1.1 Abstract

We examine the effect of industry makeup as a driver of disease growth rate using COVID-19 data across all US counties. For each day, we regress the COVID-19 growth rate on the percentage shares of each industry in a county's economy, and show how particular industries are regularly correlated with certain impacts on disease growth rate. After this initial regression, we add important covariates, such a lockdown indicator, and show that while the specifics of our coefficients change, the many of the previously significant industries maintain their significance. We regress on each day of a county's outbreak separately. This allows us to not only show which industries are significant, but also how their effect evolves over the course of the outbreak. The impact of these industries aligns with what one may expect – industries that rely on manual labor and close quarters contact have higher growth rates on average, whereas "stay at home" compatible industries, such as tech, and industries that have very small concentrations of employees working in a large, outdoor space, such as oil extraction, tend to have lower than average growth rates. We then discuss possible violations of our model assumptions, the varying effects of those violations, and potential extensions to deal with those effects.

## 1.2 Background

In late 2019, a novel strain of the coronavirus plagued Wuhan, China. This was not the first novel virus of the 21st century, and most have been well contained. However, in just a short six months, nearly the entire world went on lockdown. The virus has particularly plagued the United States, which has the most total cases and deaths at the time of writing.

Early on in the spread of the disease, one of the central questions that governments around the world faced was how to allocate limited medical resources in case of a full-blown outbreak in their country. For the United States, this often took the form of states deciding how to distribute things like tests, ventilators, and other resources between counties and individual hospitals. Because purchases of the relevant supplies skyrocketed as the disease spread and their production rate was relatively limited, it took a considerable amount of time for states to be delivered enough supplies to steadily deal with the spread of the disease. For states' pre-outbreak planning, the consequence of this was that the *rate of growth* of the disease in an area was just as important than the final projected number of sick. The primary supply shortage would be at the beginning of the outbreak, so an essential question was how fast the disease would spread in a region once an outbreak had started. Areas with rapid growth rates would tap into the limited supplies, while areas with slow growth rates could hold out until additional resources arrived.

One approach is to use traditional epidemiological models, but many suffer from an overreliance on data invisible to the pre-outbreak planner. In the simplest models, the effective reproductive rate of a disease $R$ is modeled by $R = R_0 x$, where $R_0$ is a basic reproduction number unique to a disease, and $x$ is the percentage of the population that is susceptible to the disease. $R_0$ is fully dependent on knowledge of the particular disease – information that is not known early in a pandemic. However, $x$ is in part dependent on population information such as age, health, and possible spread vectors, which we can roughly approximate using those known parameters.

The above motivates the following question: in the context of pre-outbreak planning, that is, using only data available pre-outbreak, what factors can you use to estimate the relative growth rate of the disease in a particular region? Furthermore, how can these factors be used as inputs into traditional epidemiological models to reduce their reliance on intra-outbreak information? We explore industry makeup as as a factor correlated with disease growth, and the significance of our results suggests the possibility of using industry makeup in epidemiological models as an estimator for pathogen spread characteristics.

# 2 Data

## 2.1 Disease Data

Our COVID-19 data comes from the New York Times (NYT)[1]. It is a compilation of total (historical) case and death counts, taken daily for each US county. This data is widely used and highly scrutinized. However, it suffers from the same flaw as all COVID-19 data: inadequate and inaccurate testing means the reported case count is always an underestimate of the true infected count.

It is worth noting that this data is not exactly identical to reports coming directly from states and counties. Rather, the NYT data set is somewhat edited. Most significantly, the NYT data set classifies cases and deaths by location of treatment and location of death, respectively, while many counties classify cases and deaths by location of residence. That being said, a major motivation for the compilation of this data set is that there are no consistent conventions between counties. We value the uniformity the NYT data provides over the direct reporting from the counties.

While the data is generally clean, there are occasionally artifacts where the total number of historical cases drops in a county. This occurs if a county adjusts their case count down without specifying which prior days were incorrect. Such artifacts are moderately rare, and we assigned those days to have a zero growth rate, to avoid the impossible reduction of total cases influencing our regression.

## 2.2 Industry and Economic Data

Our industry data is sourced from the Bureau of Economic Analysis's (BEA) CAEMP25N survey. It is a yearly survey of all counties in the US containing employment counts by industry. It divides approximately 20 industry categories into several subcategories each. It also contains state and national aggregates, which we disregarded. For some industries, we chose to use the overall category, while for others we chose to use subcategories. This is because for some groups, subcategories of essentially the same work would be separated due to a legal specification that bore little relevance to the type of work being done. At other times, the subgroups would represent vastly different industries in terms of actual work being done. We chose the combination we thought best grouped together similar types of work.

The most recent year of data is from 2018, which is the data we used. Most of the data is present, although for some smaller counties data is listed as missing. These counties mostly haven't had any COVID cases to begin with, so dropping them is of little importance, because our model is on growth given that an outbreak is happening. Out of the existing data, we have relatively high confidence in its accuracy, since much of it comes from information that must be reported to the government by law. One potential weakness in this data set is that some industries have incentive to misreport to the government. Industries with large numbers of undocumented workers, for example, could have systematically under-reported values in this data set.

## 2.3 Lockdown Data

One of our specifications uses an indicator for state lockdown and stay-at-home status. This data is sourced from Aura Vision's lockdown tracker, which is scraped from CNN's lockdown tracker[3]. It is extremely unlikely that there are errors in this data, as they fully cite each states press release, and can be verified across multiple articles. However, this data does lack information on the severity of lockdown. Some states imposed harsh lockdown restrictions, while others were relatively lax. The lack of quantifiable metrics of lockdown severity weakens the usefulness of lockdown data overall, however we believe it remains an important covariate.

# 3 Models

## 3.1 Independent Variables

In our base regression, our independent variables were percent industry shares and population density. Our reasoning behind and design of these variables is as follows:

A major factor in the rate of spread of any disease is the amount of contact between susceptible people. We argue that a major factor of the rate of susceptible contact is a person's profession. Certain industries require more in person interaction, or physical contact, while other industries are more isolated. As a consequence, certain occupations cause their employees to have additional risk of susceptible contact than others. This would then influence a person's exposure chances. Furthermore, employers in industries that rely on physical proximity to do business (manufacturing, airlines, etc) would be more reluctant to implement measures to limit that proximity, because doing so would limit their productivity. There are also potential influences of company policy and workplace culture on the behavior of their employees when not at work. All this suggests that some industries could be especially correlated with the rate of disease spread among its employees. Note that this correlation could be either positive (if an industry is rigid in its proximity requirements) or negative (if an industry is especially flexible).

To explore this, we include as our primary variables of interest the percentage share of each industry by county. The percentage share indicates the relative size of a particular industry within a county by the ratio of that industry's workers to the total working population. This allows us to gauge how influential a particular industry is on limiting the spread of coronavirus in the population of interest. To avoid our industry shares from being linearly dependent, we drop one category, the "other" category, from the data. This ensures our list of industry shares do not sum to 1.

We also include the population density in each county. We control for population density because regardless of industry, the size of a population affects the chance of contact outside of the workplace; for example, when performing universally necessary functions such as buying groceries. It also may account for any potential "critical mass" thresholds, as well as to mitigate any potential geographic covariance (large, sparsely populated counties will be less likely to spread disease). Needless to say, population density of a county is one of the most notable factors contributing to the rate of growth and not controlling for it would be a massive oversight.

In our second regression specification, we added an indicator for whether or not the current period is before or after a state-level shutdown order. Three states, Nebraska, South Dakota, and Wyoming, have a constant lockdown indicator of 0, as there was never a statewide stay-at-home order[6]. Note that while a lockdown status indicator is technically not availible pre-lockdown, it is a variable that is readily compatible with scenario analysis. For the state planner, our second model provides a tool to estimate the effect of a potential lockdown on industry effects over time.

## 3.2 Dependent Variables

Our dependent variables were the daily log growth rates of reported cases by county, calculated as follows:

$$\widetilde{r}_{i,t} = \log(\text{cases}_{i,t}) - \log(\text{cases}_{i,t-1}) \tag{1}$$

where $\text{cases}_{i,t}$ is the historical total reported cases in county $i$ by time $t$. Some useful properties of this variable is that each $\widetilde{r}_{i,t}$ is independent and, under a single condition, is equal to the true growth rate, unobscured by inaccuracies in the cases data. For each county, we started building growth rates from when there were 6 or more cases in that county. This ensures that our data for each county starts after the outbreak has already begun. We then extend this to the first 50 days of the outbreak, dropping all data after the 50 day mark. We are interested in the beginning of each outbreak, so this is justified. It also deals with strong data sparsity, with approximately half the counties having 50 or less days of data.

As previously stated, our premise is that different industries both place workers in more precarious conditions and tend respond with unique company policies. A core piece of this is that businesses have different reactions *over time.* So while two businesses may eventually put into place similar measures, when they do so depends on the nature of their work. This necessitates our dependent variable to track growth rates in a county *over time,* that is, indexed by both county and day. We arrive at this specific calculation of daily growth by county as follows:

A common (although often challenged[4]) assumption in epidemiology is that in the initial growth phase of an outbreak follows an exponential growth pattern[2]. Given initial infected $I_0$, the number of infected at time $t$ is approximately:

$$I_t = I_0 e^{rt} \tag{2}$$

for some growth parameter $r > 0$. If $I_0$ and $I_t$ are known, then:

$$r = \frac{\log(I_t) - \log(I_0)}{t} \tag{3}$$

This simple model would suggest that as time progressed, the number infected would tend to infinity. Clearly, this is not the case in real life. In actuality, disease growth is not only a function of the number of people who have it, but also the number of people who can get it, as well as the number who are immune (or recovered) from it. Even independently from our claim of industries causing different growth rates over time, this suggests a time dependent growth rate $r(I_t, S_t, R_t)$, dependent on infected count $I_t$, susceptible count $S_t$ and immune/recovered amount $R_t$.

Given that a time dependent growth rate is theoretically justified independently from our hypothesis, the question becomes how to calculate it. Assuming that the ratio of infected to susceptible is still close to 0, we can continue using the exponential model. A characteristic of the exponential growth function is that, if we take $t = 0, 1, 2, \ldots$, then:

$$\begin{aligned} I_t &= I_0 e^{rt} \\ &= I_0 e^{r(t-1)} e^r \\ &= I_{t-1} e^r \end{aligned}$$

Applying our time dependent growth rate:

$$I_t = I_{t-1} e^{r_t} \tag{4}$$

Which gives us an easy calculation to get $r_t$ for all $t$, as well as an estimate of $r_t$ using the cases data:

$$r_{t,i} = \log(I_{t,i}) - \log(I_{t-1,i}) \tag{5}$$

$$\widetilde{r}_{t,i} = \log(\text{cases}_{t,i}) - \log(\text{cases}_{t-1,i}) \tag{6}$$

This gives us an $r_{t,i}$ that is independent from every other $r_{t,i}$. Another useful property of this is that under a single condition, $r_{t,i} = \widetilde{r}_{t,i}$. This is because the growth rate is a function of the true infected count, not the known cases count, so estimating the change in log cases can ignore inaccuracies. Let $\alpha_t$ be the fraction of the infected population that is known in period $t$ (that is, let $\alpha_t = \frac{\text{cases}_t}{I_t}$). If $\alpha_t = \alpha_{t-1}$, then we get the following equality:

$$\begin{aligned} \widetilde{r}_t &= \log(\text{cases}_t) - \log(\text{cases}_{t-1}) \\ &= \log(\alpha_t I_t) - \log(\alpha_{t-1} I_{t-1}) \\ &= \log(I_t) + \log(\alpha_t) - \log(I_{t-1}) - \log(\alpha_{t-1}) \\ &= \log(I_t) - \log(I_{t-1}) \\ &= r_t \end{aligned}$$

So, if $\alpha_{t,i}$ is uncorrelated with county $i$ and time $t$, then we can take $\widetilde{r}_{t,i}$ as the true growth rate. This is a strong assumption and its plausibility can be improved by adding state and county fixed effects.

Much of the above is central to the development of the well known SIR model[7], which uses differential equations to model growth over time. Unfortunately, intrinsic in this type of simulation is the relationship between infection, susceptibility, and recovery. That is, it is impossible to build an SIR model (and others like it) without quantifiable facts about how infectious a disease is, as well as how long one stays sick with it. For the pre-pandemic state planner with no accurate disease information, such a model requires parameters unavailable to them. This guides us to our regression based model.

### 3.3 Regression Specification

To calculate how industries affect spread at each stage of the outbreak, we regress our daily growth data on our county data, separately for each day.

$$r_t = \beta_t X + \omega_t p + \theta_t \chi_t + \varepsilon \tag{7}$$

where $r_t$ is the vector of growth rates for all counties at time $t$, $X$ is the matrix of industry shares for all counties, $p$ is the vector of population densities for all counties, and $\chi_t$ is the vector of lockdown indicators at time $t$. In the simplified model, $\chi_t$ is removed.

For 50 days of growth data, this gives us 50 separate regressions, each showing the effect of industries at a particular point in the timeline of the outbreak. For example, regression 1 shows the effect of each industry on the first day of the outbreak, regression 2 shows the effect on the second day, and so on. Aggregating these estimates, we can build a model of how the effects evolve over time. The first model (1) is with only industry data and population density. The second model (2) is with industry data, population density, and a lockdown status indicator.

## 4 Results

The regression specifications outlined above was run on 50 days of growth rates with a case minimum of $n = 6$. The maximum number of growth rates for a single county under these parameters is 98; that county being King County, WA, the starting point of the outbreak in the United States. The value of $n$ is chosen to accommodate both for initial testing lag and to let the outbreak develop into the exponential growth stage. $n$ was chosen before any regressions were ran, and was selected after analysis for a reasonable time delay that also did

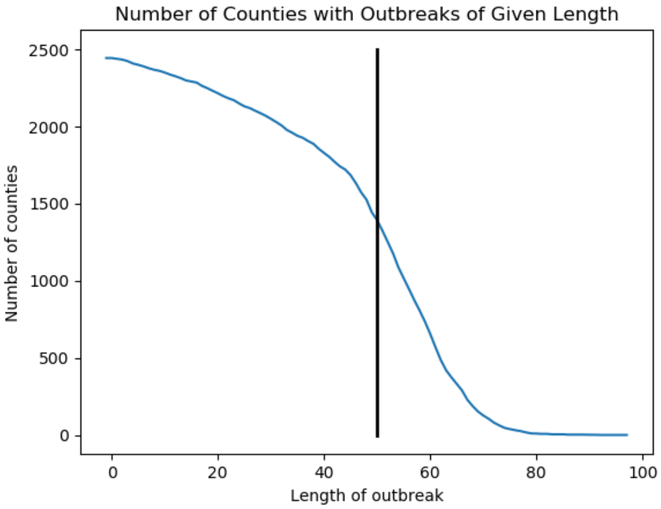

Figure 1: Cumulative distribution of the number of counties with minimum outbreak length $\ell$. We see that $\ell = 50$ (black line) gives us access to a long time interval while preserving the data of over half the infected counties in the United States.

not remove too many counties from the series of regressions due to a lack of data (see Figures 1 and 2). Significance of a coefficient was determined by a $p$-test with standard significance level $\alpha = 0.05$.

Our quantitative results across all the regressions are best shown in the following table, which, for each model, shows significant industries, how many days it was significant for, and how many days the coefficients passed a robustness check. The robustness check is a simple linear model reduction, where covariates that were insignificant in the original model ($p > \alpha$) were dropped. A covariate was deemed "robust" if it remained significant in the reduced model. Because we run 50 regressions, there is the issue of multiple hypothesis testing. However, in this case, we are only concerned with the several industries that had significance across a large number of days. It would be a mistake to take, say, "utilities," which had a single significant day, and use it as a justification for utilities driving COVID-19 spread (or in this case, reducing the spread rate). On the other hand, we can look at admin services, with a very frequently significant coefficient, and with more confidence we can say that a county's economic reliance on admin services has a real affect on growth throughout the outbreak.

From the below table, we see the top industries that were significant throughout the examined time periods: administrative and support and waste management and remediation services (administration), wholesale trade, transportation and warehousing, educational services, and manufacturing. We see that the other two independent variables, population

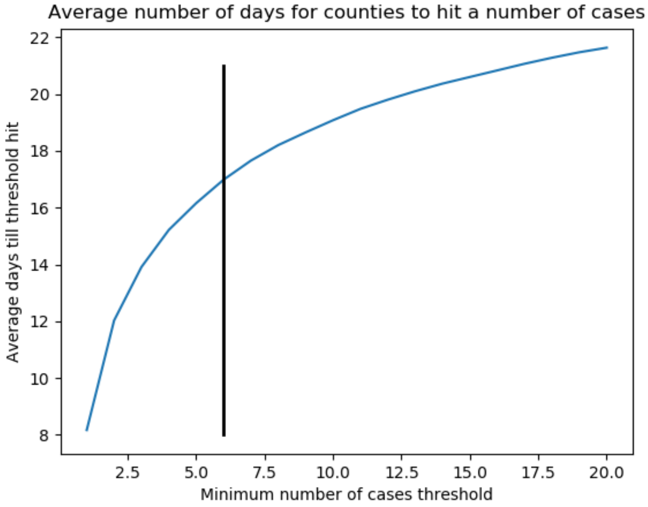

Figure 2: The average "time lag" between the first reported case and the $n$th reported case. We see that choosing $n = 6$ (black line) gives an average lag of 16 days, which balances both initial testing lag and allows for growth into the exponential growth stage

density and the lockdown indicator, are also significant for a large percentage of the examined time period. An interesting observation is the drastic reduction in significant days for educational services and a drastic increase in significant days for manufacturing when the lockdown indicator is added.

Our significant industries are indeed robust; we see a delta of 0 for our top 4 significant industries in both models between days significant and days robust.

Raw regression tables for all regressions and both models are included in an attached ZIP file. Two example regression tables from this collection, as well as a robustness check for our basic model, are included in the appendix.

We further nullify the risk of spurious regression results by turning our focus to the more interesting application of our regression coefficients: tracking significant effects of industries over the course of the outbreak. In Figure 3, we take our most frequently significant industries from (2) and show how their significant coefficients change over time. We see a strong tendency towards significance at the beginning of the outbreak, and diminishing rates of significance later on. This reinforces the validity of our model. Our assumption is that industries would react differently over time. Given enough time, pressures and regulations they would all be forced to change. This implies most of our effect should be seen at the beginning of the outbreaks, as different industries are still making independent decisions mostly without any outside social or governmental pressures. Figure 3, which maps the coefficients over time, reflects this reality.

| Industry | (1) num days signif. | (1) num days robust | (2) num days signif. | (2) num days robust |
|---|---|---|---|---|
| administrative and support and... | 22 | 22 | 20 | 20 |
| density | 19 | 18 | 17 | 17 |
| wholesale trade | 18 | 18 | 12 | 12 |
| transportation and warehousing | 16 | 16 | 13 | 13 |
| educational services | 10 | 10 | 4 | 4 |
| management of companies and en... | 8 | 8 | 5 | 5 |
| mining, quarrying, and oil and... | 7 | 3 | 7 | 3 |
| professional, scientific, and ... | 7 | 7 | 2 | 2 |
| real estate and rental and lea... | 7 | 7 | 3 | 3 |
| manufacturing | 6 | 4 | 10 | 7 |
| health care and social assista... | 5 | 5 | 5 | 5 |
| forestry, fishing, and related... | 4 | 3 | 4 | 3 |
| information | 4 | 4 | 2 | 2 |
| finance and insurance | 4 | 4 | 4 | 4 |
| state government | 3 | 3 | 2 | 2 |
| const | 3 | 2 | 5 | 2 |
| government and government ente... | 3 | 2 | 3 | 2 |
| federal civilian | 2 | nan | 2 | 1 |
| arts, entertainment, and recre... | 2 | 2 | 2 | 2 |
| accommodation and food service... | 2 | nan | 2 | 1 |
| military | 2 | 1 | 3 | 1 |
| retail trade | 2 | 2 | 2 | 2 |
| construction | 1 | nan | 1 | 1 |
| state and local | 1 | nan | 2 | nan |
| utilities | 1 | nan | 1 | nan |
| lockdown_delta | nan | nan | 21 | 21 |

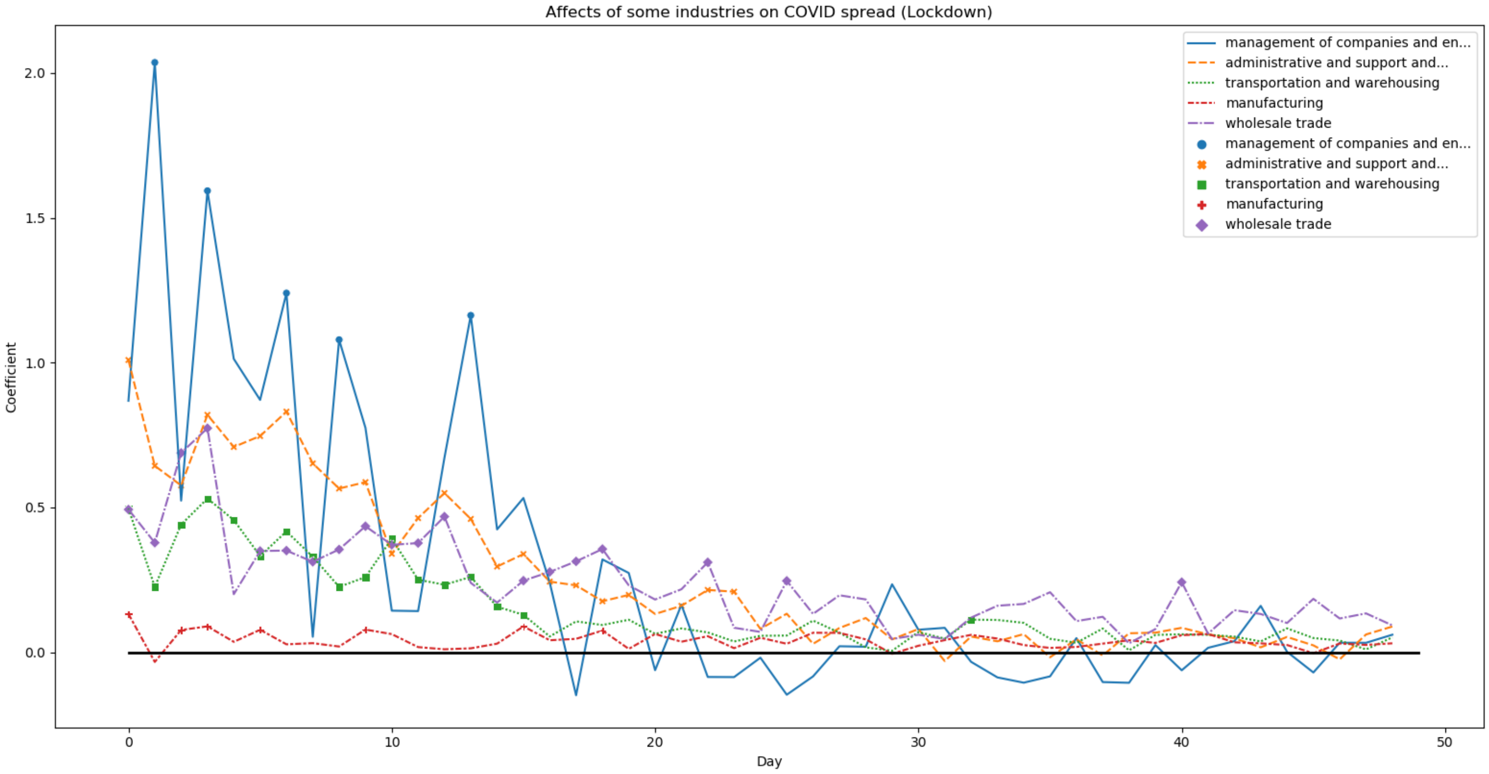

Figure 3: Industry affect on growth over time, for most frequently significant industries. Points mark significant days. We can see that most days are significant towards the beginning of the outbreak, which aligns with our expectations.

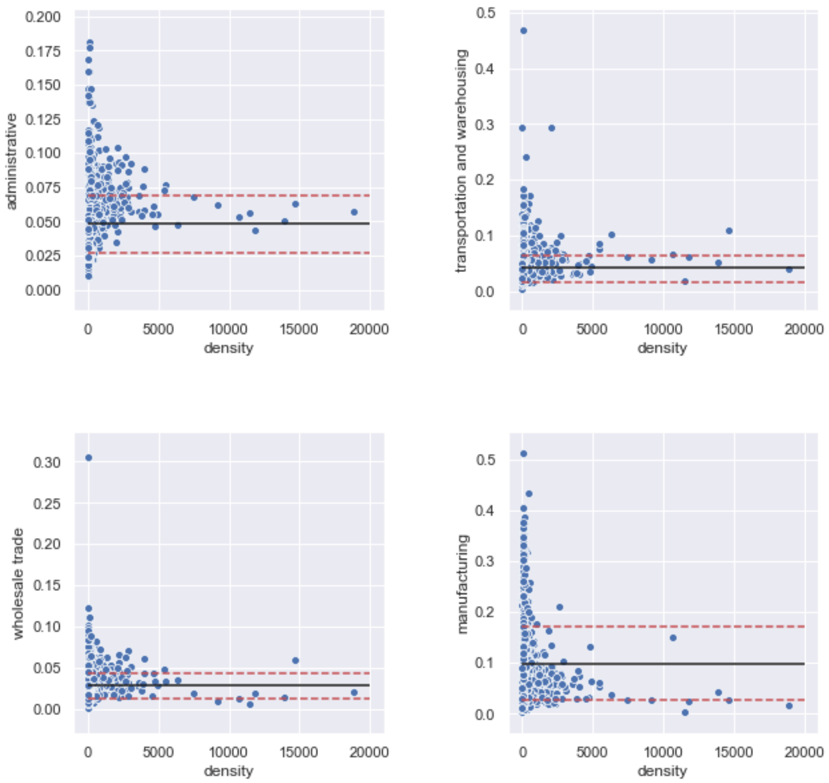

Figure 4: We plot shares of the significant industry over population density. We see that, with the exception of manufacturing, the share of industry at higher densities remain relatively clustered around the mean. Black line is the mean, and the red dashed lines are 1 standard deviation from the mean.

We also consider some implications of the significant industries identified in our model; administration, transportation and warehousing, wholesale trade, and manufacturing. As we assumed in our hypothesis, these industries heavily involve movement and person-to-person interaction, influencing the susceptible contact risk. A concern that might arise is that the industry share and the population density is correlated. After all, it would be reasonable to imply that a densely populated city will require a more robust public transportation network than a sparsely populated one. However, Figure 4 shows that, aside from manufacturing, which has industry shares dropping notably below the mean at high densities, there is no notable correlation, at least visually, between density and our respective industries. We are hesitant to apply a Pearson's correlation coefficient test because, as shown in Figure 5, our shares data is not normally distributed. Nonetheless, from these results it stands to reason that these significant industries should be used to help planners distinguish higher-risk counties by the shares of these industries.

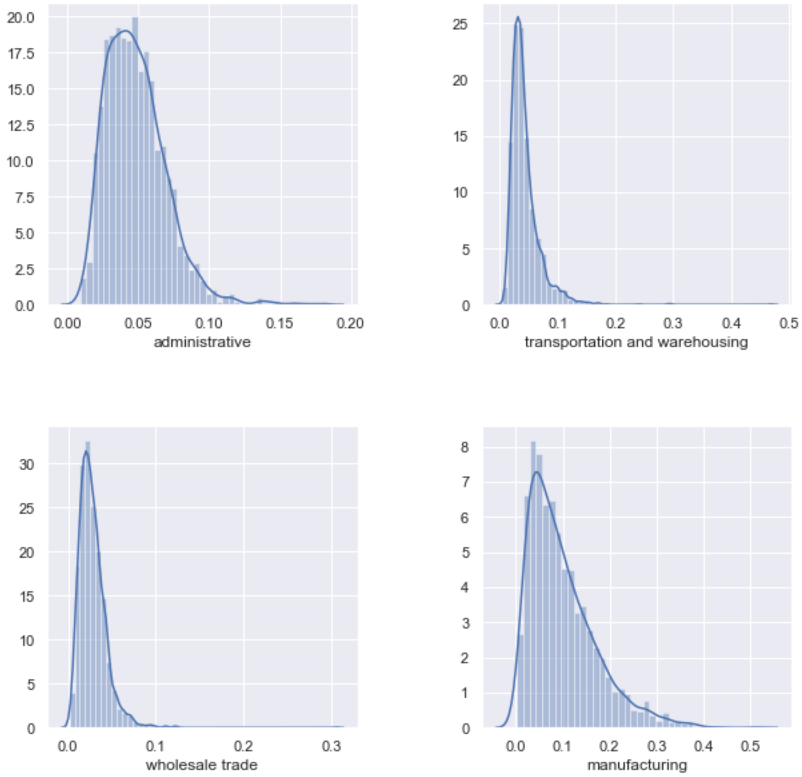

Figure 5: Distribution of industry shares.

## 4.1 Violations of Assumptions

There are several potential points of failure in this model that make it very permitting to possible extensions.

The first lies with the COVID-19 data. An inescapable issue is testing accuracy. Our definition of the rate of growth allows us to ignore testing accuracy only under the assumption that inaccuracy is uncorrelated with county and time. This assumption can easily fail. As time progresses, wider testing availability will increased the ratio of identified cases to total cases, so there is time dependence. However, this is not necessarily a fatal flaw in our model. The time frame employed takes the data from the beginning of the outbreaks, thus most cases appeared far before testing caught up to demand. Time dependency occurs at the end of our data. This can be solved by including data of testing capability over time for each county. The analysis in section 3.2 shows that time correlation would reduce the significance of our results. Correlation with county would have an uncertain affect.

Using deaths data to compensate for this is a solution in theory, because deaths are not a function of testing rates. In practice, this is difficult to implement. The primary concern is that death data is far too sparse to be useful, with the vast majority of counties having no more than a handful of deaths, if any at all. Most deaths are concentrated in large urban centers, which would then also bias the results towards large, dense counties. We see in Figure 4 that the most variance of industry shares lies in small to moderately populated counties. A secondary concern is that deaths data itself has measurement issues, just like infection data. Many deaths caused by COVID-19 are not reported as such. A proposed solution to this is to calculate "excess deaths," the number of deaths in a time period above what is expected without COVID-19. However, this does not work for our model because it will include the deaths due to hospital overflow, logistical breakdowns, and economic strain. These deaths are COVID-19 related but not caused by the virus itself, and thus doesn't indicate the rate of growth. The combination of both measurement issues and sparsity issues is what drove us to ultimately measure our dependent variable on deaths, not cases, for this paper. Future research, when deaths data becomes more replete, would be an strong way to validate the results presented here.

There is also little guarantee of independence between testing accuracy and industry. Some sectors require perfect health as a condition of employment. Workers in these industries then have incentive to get tested only for severe cases. This means that our true number sick varies by industry, and the testing inaccuracy is then correlated with industry makeup. Despite this flaw, it should not change the results. If anything, if we had accurate data then it would increase the significance of our results, as we would see a large increase in the sickness rate associated with certain industries. For this reason, while the dependence of testing accuracy on industry is certainly biasing, it would only strengthen our results if it were not a factor.

Another potential issue is that our model assumes no movement between counties. The NYT data documentation[1], there were many cases in which an individual fell sick in one county, and then was treated in a hospital in another. Most counties register this case in the county of residence of the patient, however NYT registers this in the county of treatment. For our model, we would ideally have both pieces of information. In the period of infection, the individual should be counted in their original county's growth rate, but for subsequent periods, they need to be counted in the county in which they are treated. This is not a critical issue. Most people who are not treated in their home county move to a facility with more resources to treat COVID-19. This means that the true growth rate of those unprepared counties is higher than reported by NYT, making the "unsafe" industries common in those locations even more unsafe. This would imply underestimation of the growth increase of "unsafe" industries, and overestimation of the growth decrease of "safe" industries, thereby increasing the true significance of "unsafe" industry results. The issue can also be addressed in future investigation by the inclusion of a mobility indicator or a mobility index that describes the amount of movement within a county.

A final flaw is inconsistencies in county reporting. As commented on by NYT[1], there are times when total case numbers drop in a county due to post-facto adjustments, such as transferring cases between counties. This both induces negative growth rates in one county and causes an excess positive growth rate in another county. While these negative growth rates only occur a handful of times, we are not able to identify points in the data in which the negative growth rate is masked by a corresponding positive growth rate. For example, if a county reassigned 10 cases to another county but also added 20 more in that period, then we only see the net 10 gain. Even with a strategy to properly handle negative growths, the negative adjustment in this case is a hidden effect.

Our industry data is far more reliable. It stands to note is that there are cases when industry data gathered by the government is inaccurate. One case is industries that hire a disproportionate amount of undocumented workers, which may lead to the reporting of lower employment figures. In this example, undocumented workers would mostly be involved in industries centered around manual labor. This means that the true effect of manual labor industries will be lower than our estimates, which damages the internal validity of our conclusions with respect to those industries.

Lastly, the assumption about exponential growth in the early phases of disease outbreak has been challenged in recent years, however epidemiology is still undecided on how valid those contrary results are. A recent claim is that the exponential growth is actually the sum of a collection of sub-exponential growths that happen in isolated parts of the region in question.[4].

# 5 Conclusion

## 5.1 Relation to Existing Models

A unique trait of this problem is that while the study of disease spread and epidemiology in general is a very developed field, many standard models make use of information not necessarily available in the case of a new, unknown pathogen[5]. For example, the SIR model, standard in the study of disease spread, is heavily dependent on the number infected, the number susceptible, and the number already immune, as well as specific disease information such as how contagious it is and how long it takes to stop being infectious[2]. Even some time after an outbreak begins, information on these parameters can be highly uncertain. More generally, much of predictive analysis in epidemiology is simulation based, and simulations are inherently built around already knowing the dynamics of a system, and examining how shocks to that system change outcomes. Therefore, for prediction on a new pathogen, simulations are inextricably tied to assumptions about that pathogen's dynamics. While we do make many assumptions in our analysis, using this regression based model allows our assumptions to be tied to economics and the nature of pandemic spread in general (lack of accurate testing, government intervention, etc). This, to some extent, frees our conclusions from being particular to a certain disease, and enables potential extension work to come from economic ideas, rather than developments in COVID-19 research. While this model can hardly be used for making useful predictions, a development of it could help draw conclusions about communities vulnerable to any arbitrary pandemic, and in turn, help inform our motivating question of resource allocation pre-outbreak.

## 5.2 Discussion

In this analysis, we showed how the industry makeup of a county can have an influence on the spread of a disease, using COVID-19 as a case study. Using a daily exponential growth rate, we tracked the effect of each industry over time, and found a steady and clear trend, where some industries were far more likely to see spikes in infections early on. This supports our hypothesis: the nature of certain industries and their differing incentives can cause different outcomes in disease spread. Crucially, our first model exclusively used data available before COVID-19 was even identified. The reduced reliance on the nature of COVID-19 as a pathogen helps our model inform the problem of the state planner – how to preemptively allocate limited resources between counties according to their respective needs. Our model not only gives the planner more information on where to allocate resources at the start of the pandemic; our time dependence gives information on when the influence of those industries will become imperceptible as well. Our second model is easily compatible with scenario analysis to remove its dependence on lockdown date, and with further development could even be used to analyse the effect of lockdown itself. Finally, with additional analysis, the results of our model could be used to inform modern epidemiological simulation models, which tend to have more predictive power, but also rely on specific disease characteristics which are invisible initially.

# 6 Appendix: Select Example Tables

Day 3 Basic Model

| | coef | std err | t | P> \|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **const** | 0.0038 | 0.031 | 0.124 | 0.901 | -0.057 | 0.064 |
| **density** | 1.78e-05 | 7.96e-06 | 2.236 | 0.025 | 2.19e-06 | 3.34e-05 |
| **forestry, fishing, and related...** | -0.1254 | 0.211 | -0.595 | 0.552 | -0.539 | 0.288 |
| **mining, quarrying, and oil and...** | 0.0379 | 0.181 | 0.209 | 0.834 | -0.318 | 0.393 |
| **utilities** | -0.6523 | 0.787 | -0.829 | 0.407 | -2.195 | 0.891 |
| **construction** | -0.2052 | 0.117 | -1.759 | 0.079 | -0.434 | 0.024 |
| **manufacturing** | 0.0767 | 0.057 | 1.340 | 0.180 | -0.036 | 0.189 |
| **wholesale trade** | 0.6882 | 0.218 | 3.164 | 0.002 | 0.262 | 1.115 |
| **retail trade** | 0.1705 | 0.163 | 1.045 | 0.296 | -0.150 | 0.491 |
| **transportation and warehousing** | 0.4406 | 0.115 | 3.843 | 0.000 | 0.216 | 0.665 |
| **information** | -0.0794 | 0.609 | -0.130 | 0.896 | -1.274 | 1.115 |
| **finance and insurance** | 0.0500 | 0.234 | 0.213 | 0.831 | -0.409 | 0.509 |
| **real estate and rental and lea...** | 0.6961 | 0.255 | 2.728 | 0.006 | 0.196 | 1.196 |
| **professional, scientific, and ...** | 0.2595 | 0.166 | 1.560 | 0.119 | -0.067 | 0.586 |
| **management of companies and en...** | 0.5233 | 0.553 | 0.946 | 0.344 | -0.561 | 1.608 |
| **administrative and support and...** | 0.5758 | 0.130 | 4.427 | 0.000 | 0.321 | 0.831 |
| **educational services** | 0.7560 | 0.346 | 2.182 | 0.029 | 0.077 | 1.435 |
| **health care and social assista...** | 0.0914 | 0.074 | 1.233 | 0.218 | -0.054 | 0.237 |
| **arts, entertainment, and recre...** | 0.0200 | 0.387 | 0.052 | 0.959 | -0.738 | 0.778 |
| **accommodation and food service...** | -0.0659 | 0.135 | -0.489 | 0.625 | -0.330 | 0.199 |
| **government and government ente...** | 0.1336 | 0.053 | 2.537 | 0.011 | 0.030 | 0.237 |
| **federal civilian** | 0.1766 | 0.142 | 1.242 | 0.214 | -0.102 | 0.455 |
| **military** | 0.0062 | 0.103 | 0.060 | 0.952 | -0.196 | 0.208 |
| **state and local** | -0.0492 | 0.070 | -0.704 | 0.482 | -0.186 | 0.088 |
| **state government** | -0.0196 | 0.108 | -0.182 | 0.856 | -0.232 | 0.192 |
| **local government** | -0.1080 | 0.078 | -1.379 | 0.168 | -0.262 | 0.046 |

Day 3 Basic Model Robustness Check

| | coef | std err | t | P> \|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **density** | 2.173e-05 | 7.78e-06 | 2.792 | 0.005 | 6.47e-06 | 3.7e-05 |
| **wholesale trade** | 0.7875 | 0.196 | 4.014 | 0.000 | 0.403 | 1.172 |
| **transportation and warehousing** | 0.4545 | 0.104 | 4.388 | 0.000 | 0.251 | 0.658 |
| **real estate and rental and lea...** | 0.8380 | 0.161 | 5.219 | 0.000 | 0.523 | 1.153 |
| **administrative and support and...** | 0.7053 | 0.125 | 5.630 | 0.000 | 0.460 | 0.951 |
| **educational services** | 1.1706 | 0.320 | 3.656 | 0.000 | 0.543 | 1.798 |
| **government and government ente...** | 0.1061 | 0.029 | 3.673 | 0.000 | 0.049 | 0.163 |

## Day 3 Lockdown Model

| | **coef** | **std err** | **t** | **P> \|t\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **const** | 0.0116 | 0.031 | 0.378 | 0.706 | -0.049 | 0.072 |
| **lockdown_delta** | 0.0622 | 0.009 | 7.086 | 0.000 | 0.045 | 0.079 |
| **density** | 1.596e-05 | 7.54e-06 | 2.116 | 0.034 | 1.17e-06 | 3.07e-05 |
| **forestry, fishing, and related...** | -0.1526 | 0.208 | -0.732 | 0.464 | -0.561 | 0.256 |
| **mining, quarrying, and oil and...** | 0.0548 | 0.180 | 0.304 | 0.761 | -0.298 | 0.408 |
| **utilities** | -0.6007 | 0.784 | -0.767 | 0.443 | -2.137 | 0.936 |
| **construction** | -0.1876 | 0.116 | -1.623 | 0.105 | -0.414 | 0.039 |
| **manufacturing** | 0.1026 | 0.057 | 1.791 | 0.073 | -0.010 | 0.215 |
| **wholesale trade** | 0.6244 | 0.217 | 2.878 | 0.004 | 0.199 | 1.050 |
| **retail trade** | 0.1450 | 0.163 | 0.890 | 0.373 | -0.174 | 0.464 |
| **transportation and warehousing** | 0.3467 | 0.114 | 3.051 | 0.002 | 0.124 | 0.569 |
| **information** | -0.1933 | 0.596 | -0.324 | 0.746 | -1.361 | 0.975 |
| **finance and insurance** | 0.0203 | 0.231 | 0.088 | 0.930 | -0.432 | 0.472 |
| **real estate and rental and lea...** | 0.5609 | 0.248 | 2.260 | 0.024 | 0.074 | 1.048 |
| **professional, scientific, and ...** | 0.1452 | 0.148 | 0.982 | 0.326 | -0.145 | 0.435 |
| **management of companies and en...** | 0.3035 | 0.534 | 0.568 | 0.570 | -0.744 | 1.351 |
| **administrative and support and...** | 0.3241 | 0.131 | 2.471 | 0.014 | 0.067 | 0.581 |
| **educational services** | 0.4891 | 0.332 | 1.475 | 0.140 | -0.161 | 1.139 |
| **health care and social assista...** | 0.0861 | 0.073 | 1.177 | 0.239 | -0.057 | 0.230 |
| **arts, entertainment, and recre...** | -0.0633 | 0.385 | -0.164 | 0.870 | -0.819 | 0.693 |
| **accommodation and food service...** | -0.0581 | 0.134 | -0.434 | 0.665 | -0.321 | 0.205 |
| **government and government ente...** | 0.1296 | 0.053 | 2.463 | 0.014 | 0.026 | 0.233 |
| **federal civilian** | 0.1971 | 0.138 | 1.428 | 0.153 | -0.074 | 0.468 |
| **military** | -0.0432 | 0.097 | -0.447 | 0.655 | -0.233 | 0.146 |
| **state and local** | -0.0242 | 0.070 | -0.347 | 0.729 | -0.161 | 0.112 |
| **state government** | -0.0703 | 0.106 | -0.661 | 0.509 | -0.279 | 0.138 |
| **local government** | -0.1059 | 0.078 | -1.358 | 0.175 | -0.259 | 0.047 |

# References

[1] https://github.com/nytimes/covid-19-data

[2] Anderson R.M., May R.M. Oxford University Press; Oxford: 1991. Infectious Diseases of Humans.

[3] https://auravision.ai/covid19-lockdown-tracker/

[4] Chowell G., Viboud C., Hyman J.M., Simonsen L. The Western Africa ebola virus disease epidemic exhibits both global exponential and local polynomial growth rates. PLoS Curr. 2015:7.

[5] Kermack, W., & McKendrick, A. (1937). Contributions to the mathematical theory of epidemics IV. Analysis of experimental epidemics of the virus disease mouse ectromelia. Journal of Hygiene, 37(2), 172-187. doi:10.1017/S0022172400034902

[6] https://www.cnn.com/2020/04/13/politics/asa-hutchison-arkansas-coronavirus/index.html.

[7] H.W. Hethcote, Qualitative analyses of communicable disease models, Math. Biosci. 28 (1976) 335–356