Working paper

The Coronavirus and the Cities:

Explaining Variations in the Onset of Infection and in the Number of Reported Cases and Deaths in U.S. Metropolitan Areas as of 27 March 2020

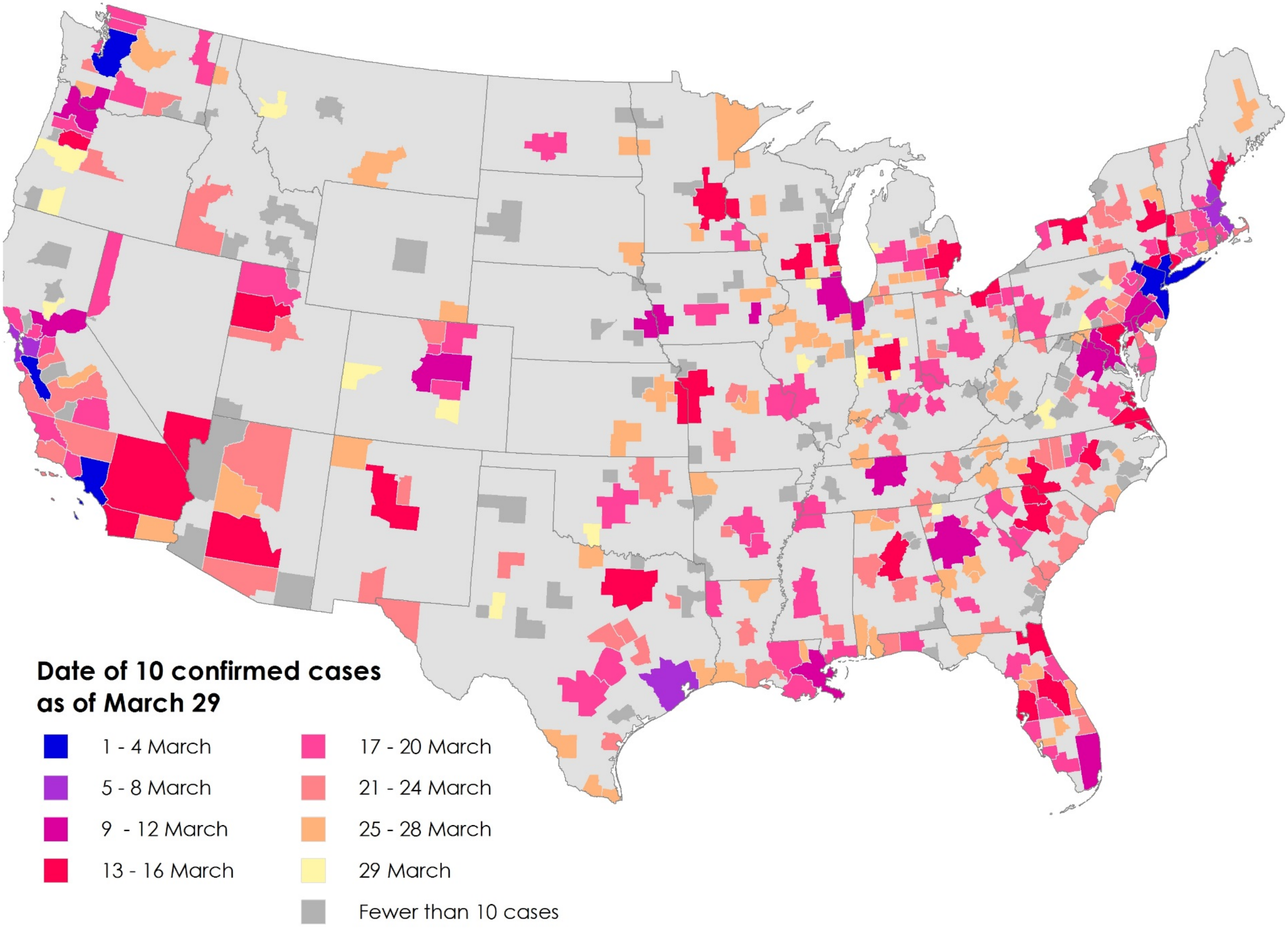
Shlomo Angel, Alejandro M. Blei, Patrick Lamson-Hall and Maria Monica Salazar Tamayo, The Marron Institute of Urban Management, New York University

31 March 2020

Press Release:

- A team of researchers led by Professor Shlomo (Solly) Angel at the Marron Institute of Urban Management at New York University has obtained new insights on the geographic spread of the Coronavirus as of 27 March 2020 by focusing on Metropolitan Areas (MSAs).
- Using data on MSAs, we sought to answer three questions: (1) Why did the onset of infection appear earlier in some cities than in others? (2) Why do some cities have more confirmed cases than others? (3) Why do some cities have more deaths than others?
- Our main findings:
- The onset of infection in a given MSA is a function of its population and its density, and—to some extent, not statistically significant—its role as a gateway to the world. Our statistical model explains 48% of the variation in the onset of infection among MSAs.
- the number of reported cases is higher in more populated and more dense metropolitan areas with more extensive testing, and with an earlier onset of infection. Our statistical model explains 81% of the variation in reported cases of infection among MSAs.
- We find that New York—like Los Angeles, San Francisco, San Jose, and Seattle—is not the *epicenter* but the *vanguard* on the pandemic front (see map). While it has by far the largest number of cases, it is not locus from which the epidemic has been spreading.
- The number of Coronavirus deaths in an MSA is a function of its population and the onset of infection in the MSA, but not of density or the share of the population above 75 years of age. Our model explains 35% of the variation in reported deaths among MSAs.

- Finally, the number of confirmed deaths can also be explained by the number of confirmed cases: a 10% increase in the number of reported infections on 27 March 2020 was associated with a 14.4% increase in the number of reported deaths on that date.
- The most important conclusion of our preliminary analysis of the Coronavirus and the cities is that variations in the geographic spread of the Coronavirus in U.S. Metropolitan Statistical Areas (MSAs) are quite predictable and explainable.


Date of 10 confirmed cases
as of March 29
1 - 4 March
5 - 8 March
9 - 12 March
13 - 16 March
17 - 20 March
21 - 24 March
25 - 28 March
29 March
Fewer than 10 cases

Executive Summary:

- A team of researchers led by Professor Shlomo (Solly) Angel at the Marron Institute of Urban Management at New York University has obtained new insights on the geographic spread of the Coronavirus as of 27 March 2020 by focusing on Metropolitan Areas (MSAs).
- The Coronavirus pandemic is, by and large, an urban pandemic: Of the total number of confirmed cases in the U.S., 96,012 or 93% were in 392 Metropolitan Statistical Areas (MSAs). It is useful, therefore, to monitor the pandemic by focusing on cities.
- The U.S. Census and the Office of Management and Budget collect data for MSAs, while data on testing for the virus is reported at the state level and data on cases of infection and death is reported at the county level. We aggregated all data by MSAs.
- Oregon and Florida report on testing at the county level. We tested the possibility of predicting the level of testing at the county and level by pro-rating state level testing data by the county share of the state population. These predictions proved reliable.
- We generated maps and tables that provide numerical and visual data at the MSA level. These maps and tables can be updated daily. Aggregating data by MSAs reveals patterns that remain hidden at the state or county level.
- For example, five MSAs have reported more deaths from the Coronavirus per 100,000 population by 27 March 2020 than New York (3.2): Albany, GA (12.8), New Orleans (7.8), Seattle (4.2), Pittsfield, MA (3.8) and Burlington, VT (3.8).
- Using data on MSAs, we sought to answer three questions: (1) Why did the onset of infection appear earlier in some cities than in others? (2) Why do some cities have more confirmed cases than others? (3) Why do some cities have more deaths than others?
- We defined the onset of infection as the number of days since 29 February 2020 by which 10 cases of infection were first reported for a given MSA (see map). We then constructed a multiple regression model to explain the onset of infection using information on MSAs.
- The first MSA to report 10 cases was the New York MSA which reported it on 1 March 2020. By 27 March, 258 MSAs—66 percent of all MSAs—reported on the onset of infection there.
- MSAs that reported on the onset of Coronavirus infection by 27 March 2020 contain 73% of the U.S. total population and a joint GDP of $16.7 trillion in 2018, accounting for 84% of the U.S. Gross Domestic Product (GDP) in that year.
- The onset of infection in a given MSA is a function of its population and its density, and—to some extent, not statistically significant—its role as a gateway to the world. Our statistical model explains 48% of the variation in the onset of infection among MSAs.

- More precisely, a 10% increase in the total population of an MSA is associated with a 1.7% decline in the number of days to the onset of infection; and a 10% increase in urban density is associated with a 1.1% decline in the number of days to the onset of infection.
- We hypothesized that the number of reported cases would be higher in more populated metropolitan areas, in more dense metropolitan areas, in metropolitan areas with more extensive testing, and in metropolitan areas with an earlier onset of infection.
- We confirmed these four hypotheses with a second multiple regression model. This model is surprisingly powerful: It explained 81% of the variation in the number of infections reported on 27 March 2020 in U.S. Metropolitan Statistical Areas (MSAs).
- More precisely, a 10% increase in the total population of an MSA is associated with a 4.6% increase in the number of reported cases of infection; and a 10% increase in density is associated with a 1.3% increase in the number of reported cases of infection.
- Furthermore, a 10% increase in the number of days since the onset of infection is associated with a 13.3% increase in the number of infections; and a 10% increase in the number of tests is associated with a 2.3% increase in reported cases of infection.
- Finally, we hypothesized that the number of Coronavirus deaths in an MSA would be a function of its population, its density, the onset of infection in the MSA and the share of the population above 75 years of age there.
- A third multiple regression model explained 35% of the variations in confirmed deaths by 27 March 2020. It confirmed that a 10% increase in the total population of an MSA is associated with an 12% increase in the number of reported deaths there.
- More importantly, the model confirmed that a 10% increase in the number of days since the onset of infection is associated with a 28.0% increase in the number of reported deaths.
- The two other variables in this model—the share of the population over 75 years of age and the share of the population living at high density have the right effect on the reported number of deaths but are not statistically significant.
- The number of confirmed deaths can also be explained by the number of confirmed cases: a 10% increase in the number of reported infections on 27 March 2020 was associated with a 14.4% increase in the number of reported deaths on that date.
- The most important conclusion of our preliminary analysis of the Coronavirus and the cities is that the geographic spread of the Coronavirus in U.S. Metropolitan Statistical Areas (MSAs) is quite predictable and explainable.

- The main reason that some MSAs report more infections than others is that the onset of infection there occurred earlier. In this sense, New York is not the *epicenter* of the pandemic but the *vanguard* on the pandemic front.
- Secondary reasons that some MSAs report more infections than others are that they are larger and denser and do more testing, but not necessarily because they contain a larger share of older people.
- All of this may be quite obvious, but in these times of uncertainty it may make provide some people some comfort to know that, for now, the onset of infection as well as the number of people infected and dying is explainable and, to an extent, even predictable.
- In subsequent analyses, we plan to update the models and possibly make them more comprehensive by including other factors in our analysis, such as per capita public health expenditures or the onset of state stay-at-home orders.

* * *

Introduction:

We can now begin to explain the geographic variations in the date of the onset of infections, in the number of confirmed cases, and in the number of deaths from the Coronavirus. Instead of focusing on states or on counties, we focus on cities, and more specifically on Metropolitan Statistical Areas (MSAs) in the United States. Others have already begun to look at the spread of the Coronavirus in U.S. cities (see, for example Cohn et al, 27 March 2020[1]). The virus does not recognize state or county boundaries and MSAs indeed cross over both county and state boundaries. MSAs are integrated urban economies with a high level of connectivity within them, suggesting that they are the appropriate units for analyzing the data on the spread of the virus.

The question that many of us are asking is ‘why do some U.S. metropolitan areas have more infections and more deaths than others?’ A number of conjectures have been advanced. The governor of the state of New York, Andrew Cuomo, for example, conjectured on 26 March that New York has more cases than any other city because it is dense and because it is an international gateway. This is a hypothesis that can now be tested with data.

[1] Cohn, N., Katz, J., Sanger-Katz, M., and Quealy, K., Some U.S. Cities Could Have Coronavirus Outbreaks Worse Than Wuhan’s, *The New York Times*, 27 March.

We have tried to answer three related questions:

- Why did some cities encounter Coronavirus infections earlier than others?
- Why do some cities have more confirmed cases of Coronavirus infections than others?
- Why do some cities have more deaths from the Coronavirus than others?

In the following sections we analyze Coronavirus data for 28 March 2020 to provide answers to each of these questions using multiple regression models. Before we can do that, we explain how we assembled the dataset for these models.

Sources of Data:

There are at total of 392 Metropolitan Statistical Areas (MSAs) in the United States and Puerto Rico. In 2018, these MSAs had a total population of 280 million people and comprised 87 percent of the population of the country. MSAs are comprised of counties, sometimes counties in different states.

We obtain data from the U.S. census and other sources on the total population of MSAs, on their 'urbanized areas', on the population density of individual census tracts within them, on their Gross Domestic Product, and on the share of the population above 75 years of age. We also obtained data on the total number of workers above 16 years of age in each MSA and the share of these workers who commute to work by public transit.[2]

In addition, we obtained data on the number of international passenger flows at the airports of each MSA in the quarter ending in June 2019.[3]

We also obtained data on the number of infections and deaths by the Coronavirus by county for 27 March 2020.[4]

[2] We obtained spatial boundary files for U.S. counties, tracts, urban areas and MSA from the US Census website, www.census.gov. The most recent year for estimates of individual MSA GDP was 2017 and this information is released by the Bureau of Economic Analysis.. The most recent census estimates of socioeconomic data for MSAs, counties, and tracts, including their populations, mode share, workers, and age, was downloaded from the website www.socialexplorer.com.

[3] International passenger flow data is associated with 315 airports in the United States. International passenger data was cross tabulated using an airport's three letter IATA code. This information is contained in the quarterly report *US International Air Passenger and Freight Statistics for June 2019*. IATA's were geocoded and matched to MSAs. A very small number of international passenger traffic arrives or departs from airports outside of any MSA.

[4] County level data for confirmed cases and deaths was obtained from the *New York Times* Covid-19 Github web page. We aggregated this county-level data to generate MSA values.

Data on testing by county is not yet available. It is available by state. We were able to obtain data on the number of tests by county in two states, Oregon and Florida. We tested the hypothesis that we could predict the number of tests in counties in these two states by pro-rating them by the population of each county. This assumes that tests are evenly distributed among the populations of states. The estimates obtained in this manner were very good. The regression line estimate in both states had an R^2 values of 0.91 and 0.87 respectively. Figure 1 below show the predicted value on the Y-axis and the actual value on the X-axis.

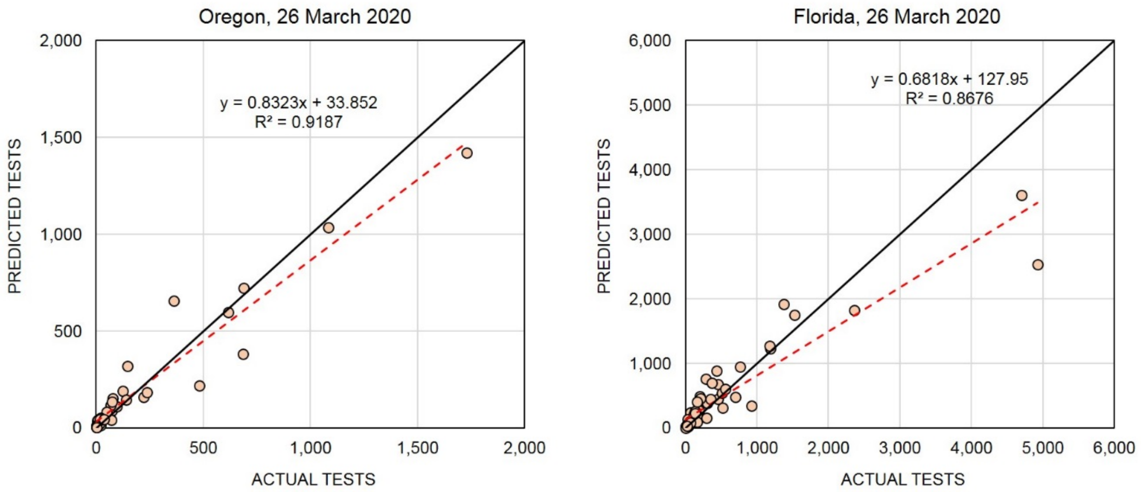


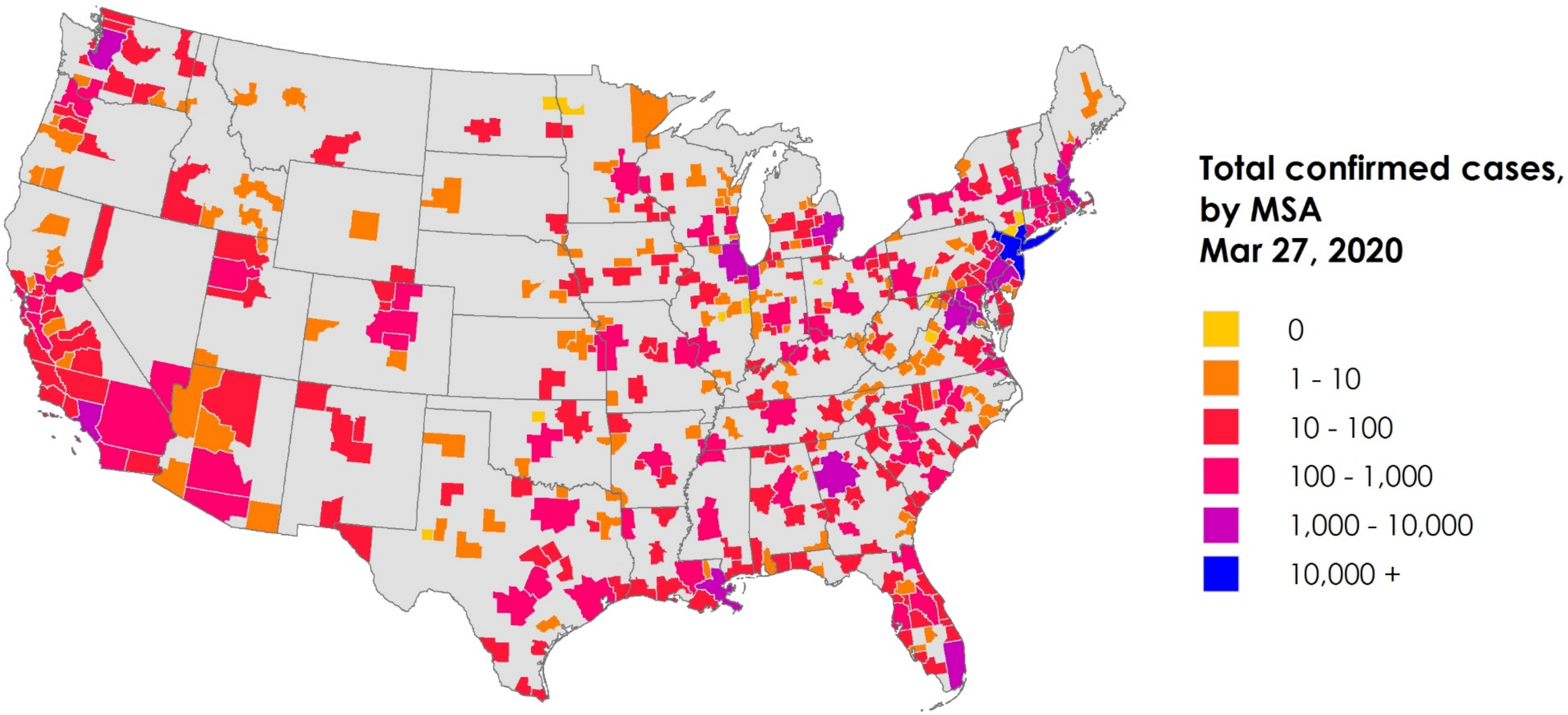
Figure 1: Estimating county level infection rates from observed rates in Oregon (26 March 2020) and Florida (27 March 2020). The X-axis gives the actual number of tests in each county. The Y-axis gives the predicted value assuming testing is evenly distributed in the state.

We used this finding to allocate the available statewide data on testing among counties. This, in turn, allowed us to estimate the number of tests in each of the MSAs on 25 March 2020.

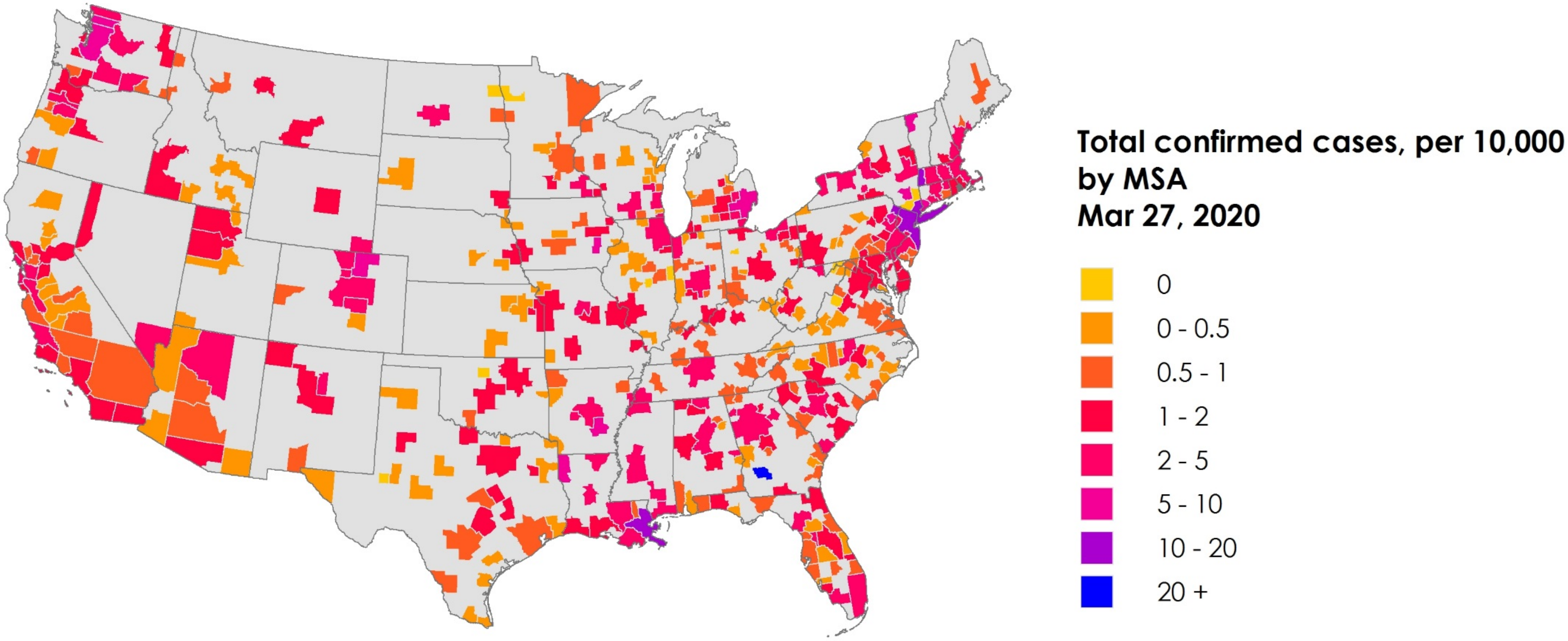
These data were used to test a number of hypotheses regarding (1) the onset of infection of the Coronavirus, (2) the number of confirmed cases and (3) the number of deaths in each MSA in the United States on 27 March 2020. This, of course, is an initial attempt at testing these hypotheses and we plan to improve on it as data becomes more plentiful.

Maps

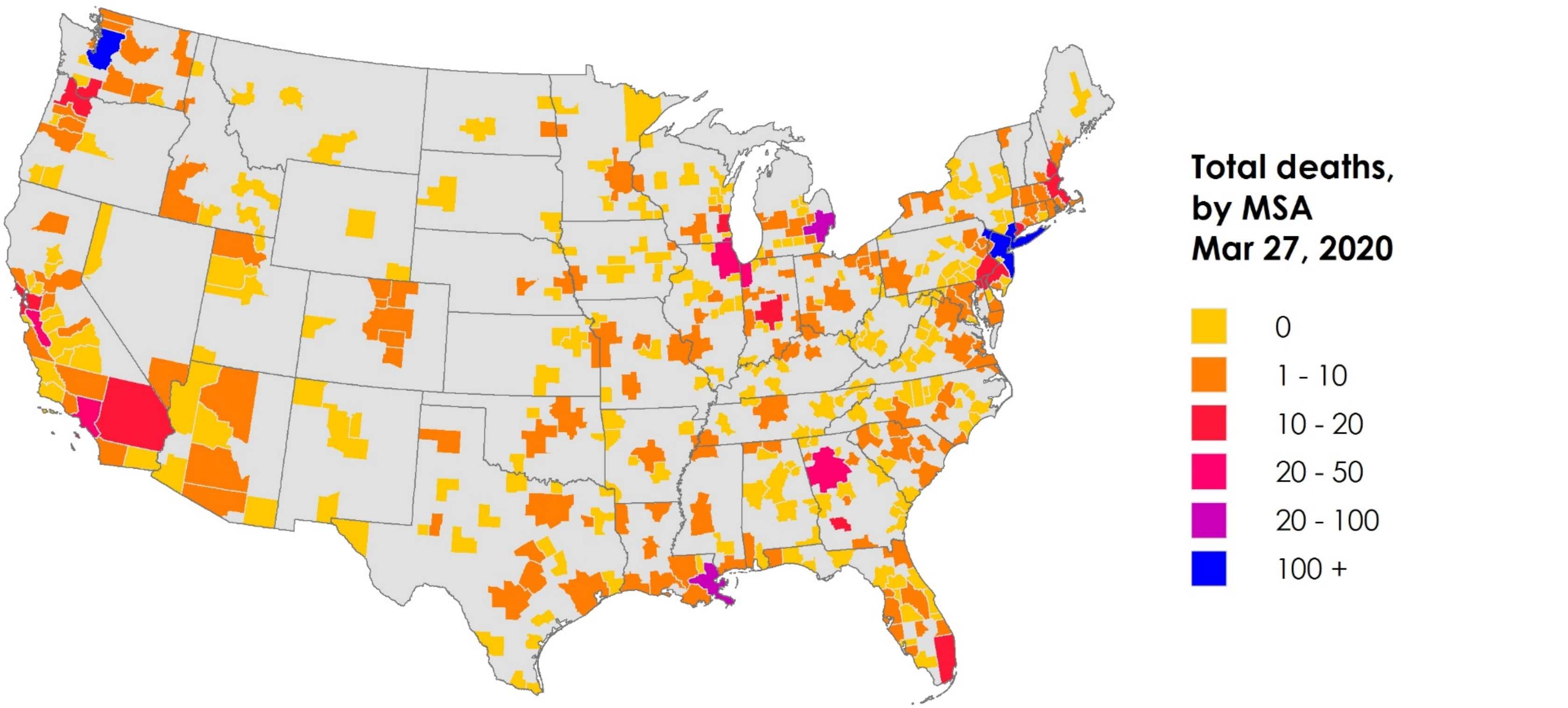
The following maps provide an overview of the spatial distribution of confirmed cases, deaths, and estimated testing rates at the MSA level. We compare the gross measure with a per capita measure, reported as per 10,000 population. Subsequent maps show the population that lives above a density threshold of 10,000 persons per mi^2 in each MSA, as well as the international passenger flows associated with airports located within MSA boundaries.

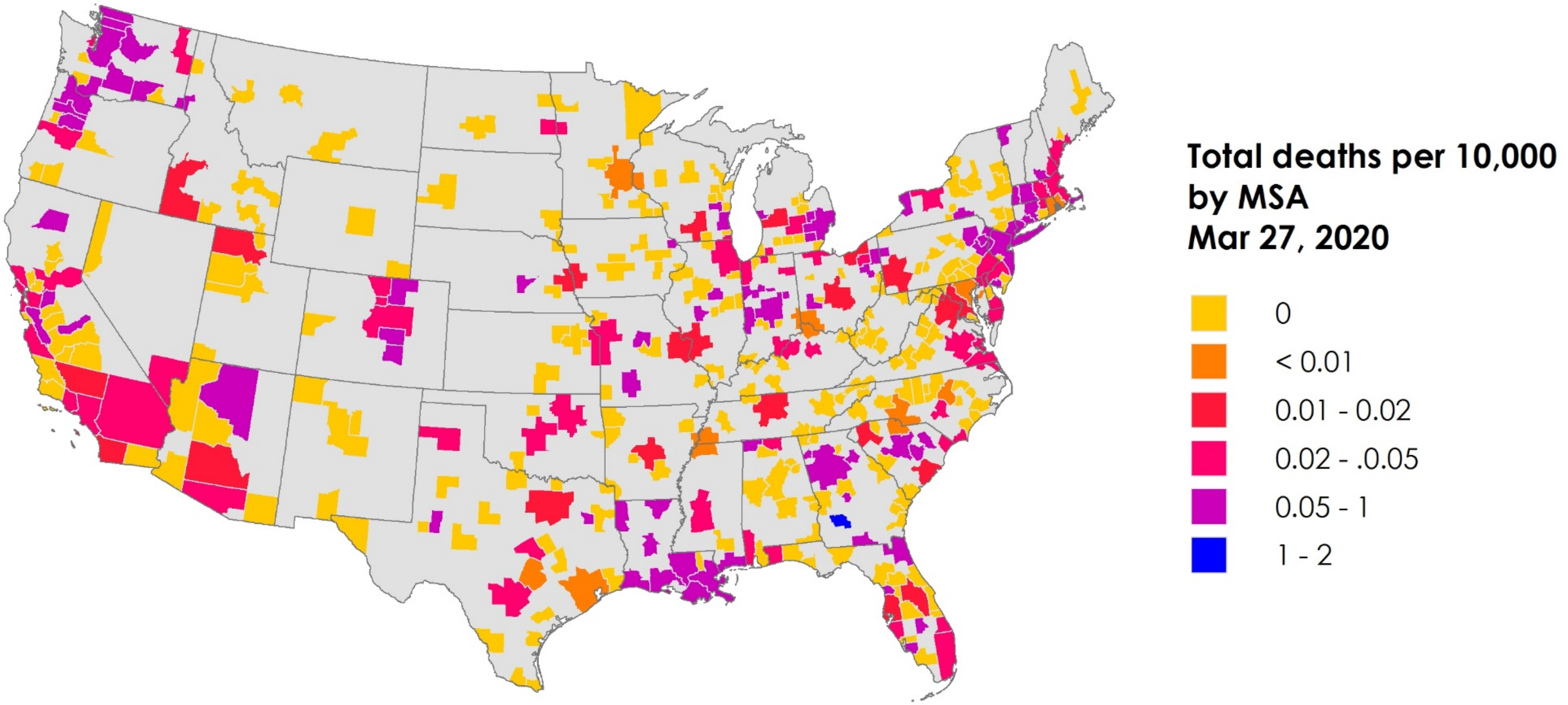


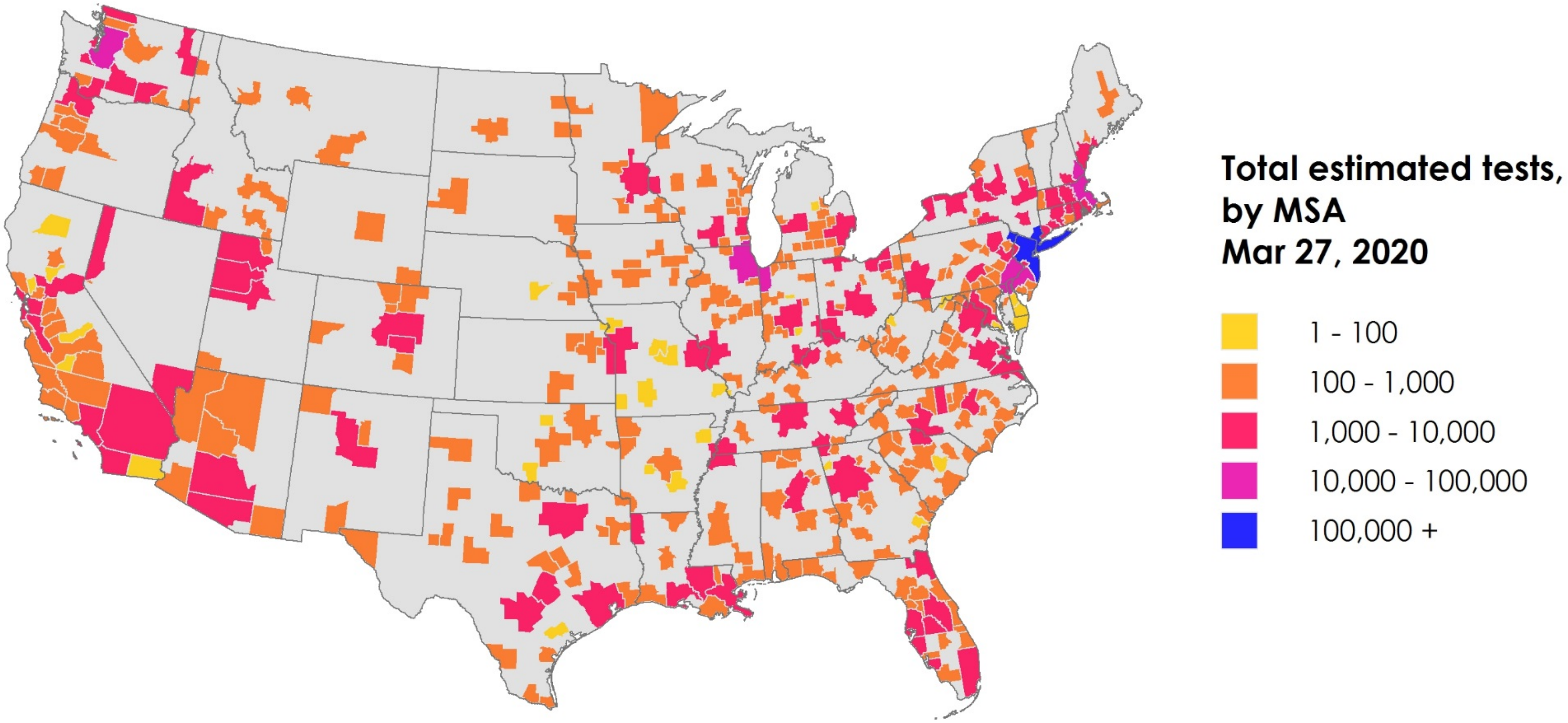
Total confirmed cases, by MSA Mar 27, 2020



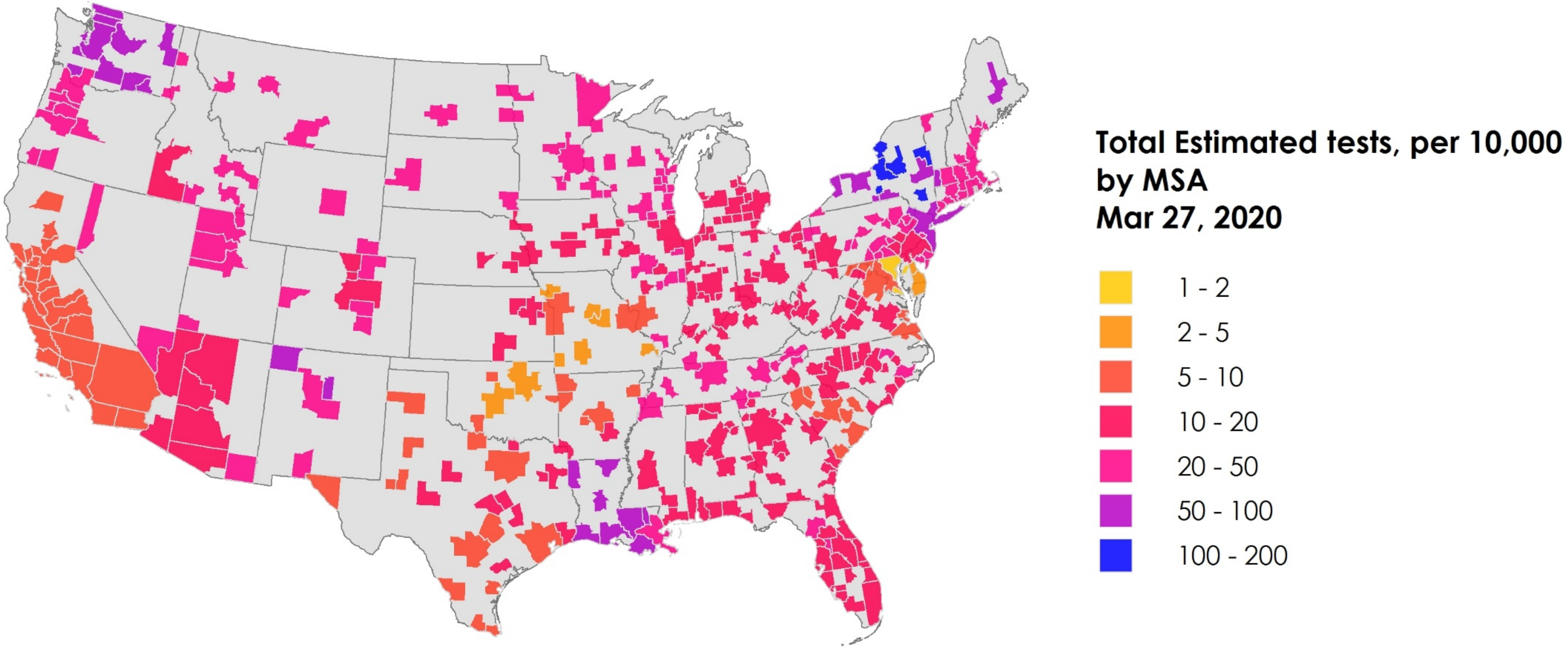
Total confirmed cases, per 10,000 by MSA Mar 27, 2020


Total deaths,
by MSA
Mar 27, 2020
0
1 - 10
10 - 20
20 - 50
20 - 100
100 +

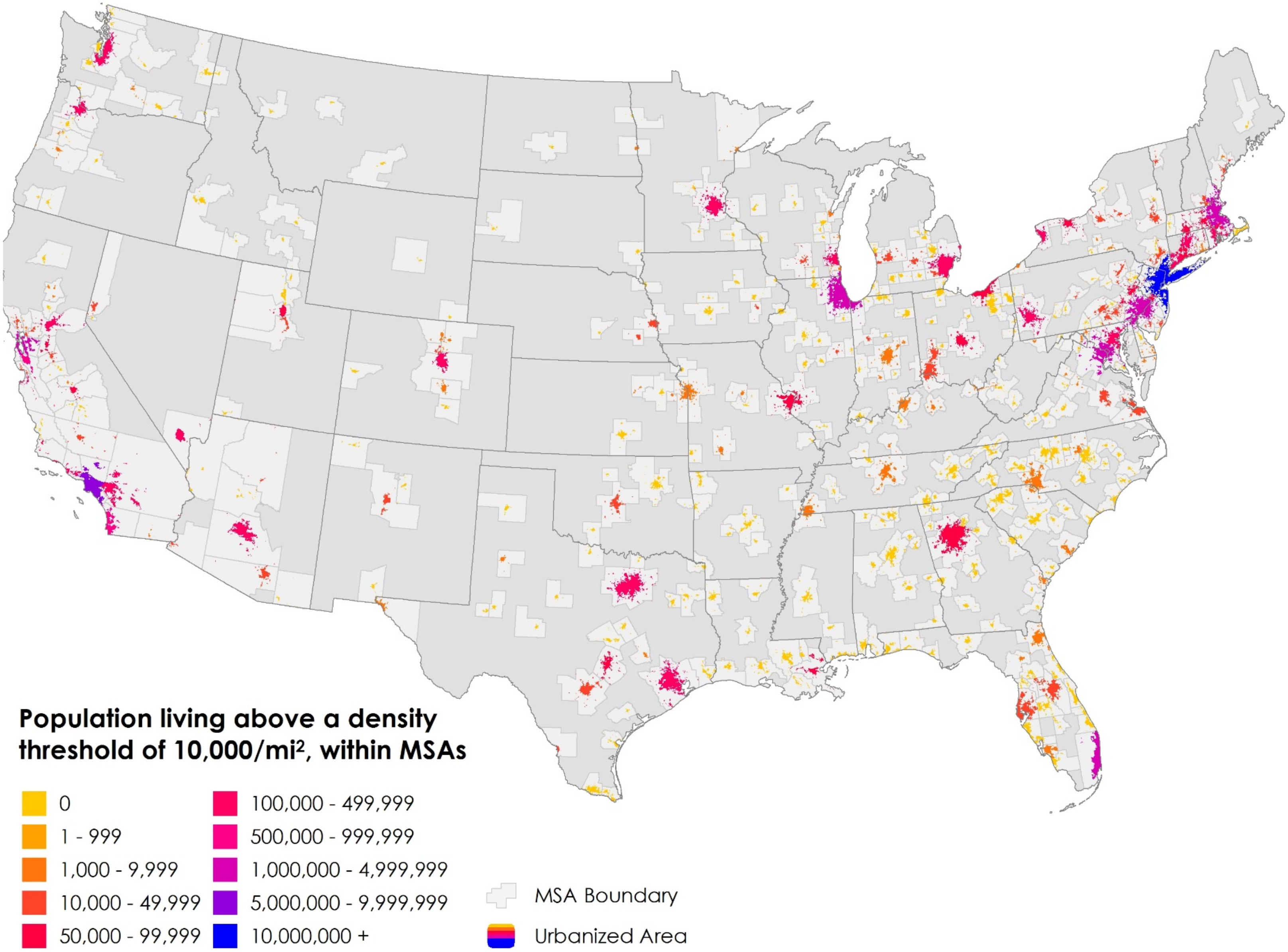

Total deaths per 10,000
by MSA
Mar 27, 2020
0
< 0.01
0.01 - 0.02
0.02 - .0.05
0.05 - 1
1 - 2



Total estimated tests, by MSA Mar 27, 2020

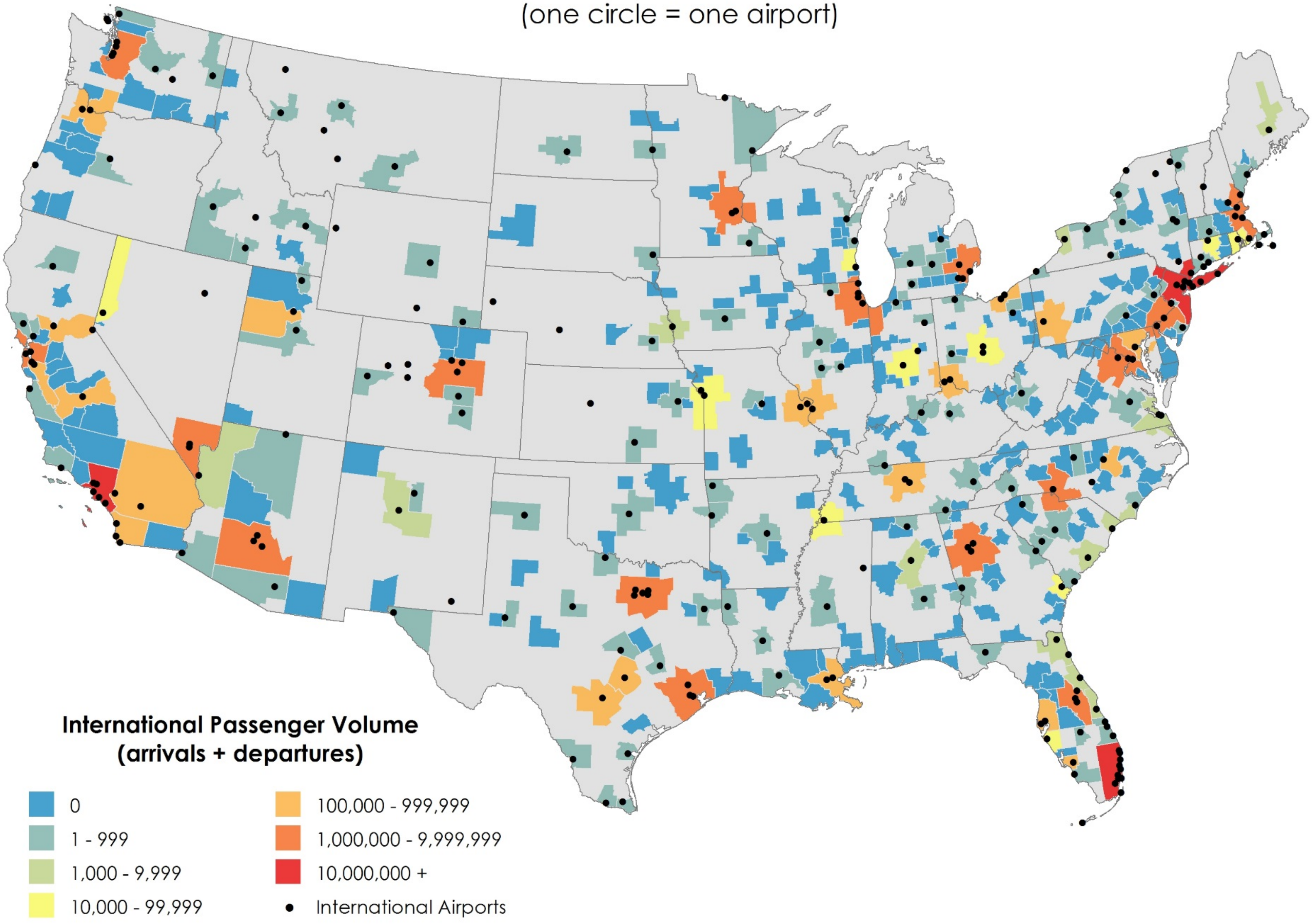


Total Estimated tests, per 10,000 by MSA Mar 27, 2020



Population living above a density threshold of 10,000/mi^2, within MSAs

MSA International Passenger Flows by Airports within MSAs, Jan. - Jun. 2019
(one circle = one airport)



Explaining the Onset of Infections in U.S. Metropolitan Areas:

Because data on the first case detecting in each MSA is sketchy, we have focused on the date by which *ten* cases were reported as the date of the onset of the Coronavirus infection in a given MSA. The specific variable that we examine as the dependent variable in our multiple regression model is the number of days after the 29th of February 2020. The first MSA to report 10 cases was the New York MSA which reported it on 1 March 2020. By 27 March, 258 MSAs—66 percent of all MSAs, containing 73% of the U.S. total population[5] and a joint GDP of $16.7 trillion in 2018, accounting for 84% of the U.S. Gross Domestic Product (GDP) in that year[6]—reported on the onset of infection there.

As Governor Cuomo conjectured, the onset of infection in New York was early because New York is a gateway city. It is also very large and very dense in comparison to other MSAs. These conjectures are indeed true. We can state with confidence that the onset of infection of the Coronavirus was earlier in larger cities and in denser cities but not necessarily in gateway cities. Indeed, three variables—the total population of the MSA, the number on international passengers per capita arriving or departing from the MSA during April-June 2019, and the share of commuters in the MSA using public transit (a proxy for high density)—explains 48 percent of the variation in the date of onset of the Coronavirus infection, measured by the number of days since 29 February 2020 that an MSA reported at least ten cases of Coronavirus infections by 27 March 2020. The multiple regression model describing these findings is shown in table 1 below.

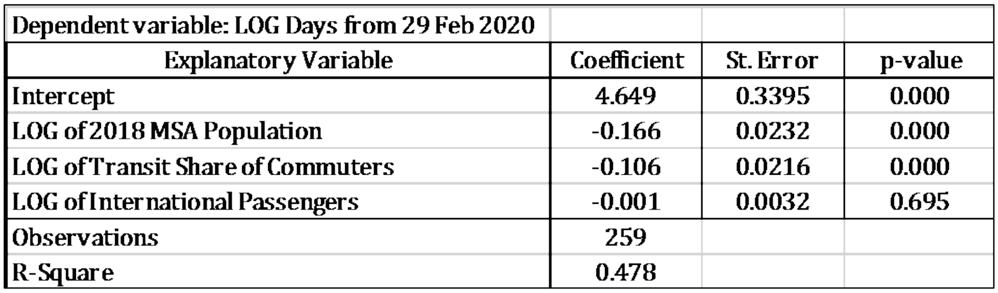
Table 1 is displayed in logarithmic form. It explains 48% of the variation in the date of the onset of Coronavirus infections in MSAs. It shows that two of the three explanatory variables postulated to explain the onset of infection in MSAs were significant at the 95% confidence levels. The logarithmic form of the model allows us to describe the coefficient of each of the explanatory variable as an elasticity: (1) a 10% increase in the total population of an MSA is associated with a 1.7% decline in the number of days to the onset of infection; and (2) a 10% increase in high-density living (measured by the share of commuters in the MSA using transit) is associated with a 1.1% decline in the number of days to the onset of infection. The third explanatory variable—the number of international passengers—has the correct effect on the onset of infections but is not statistically significant.

The key finding reported here is that the onset of infection in a given MSA is predictable. It is a function of its population and its density, and—to some extent, not statistically significant—its role as a gateway to the world. This helps explain why the onset of infections by the Coronavirus in New York, for example, was 25 days earlier than that of Fairbanks,

[5] U.S. Census, 2018 data.

[6] U.S. Bureau of Economic Analysis, 2018 data.

Arkansas. A similar statistical model, using the share of the population living above a tract density of 10,000 persons per square mile instead of transit share—yields similar results.



Dependent variable: LOG Days from 29 Feb 2020			
Explanatory Variable	Coefficient	St. Error	p-value
Intercept	4.649	0.3395	0.000
LOG of 2018 MSA Population	-0.166	0.0232	0.000
LOG of Transit Share of Commuters	-0.106	0.0216	0.000
LOG of International Passengers	-0.001	0.0032	0.695
Observations	259		
R-Square	0.478		

Table 1: Multiple regression model explaining the variation in the onset of at least 10 Coronavirus infections in U.S. MSAs as of 27 March 2020.

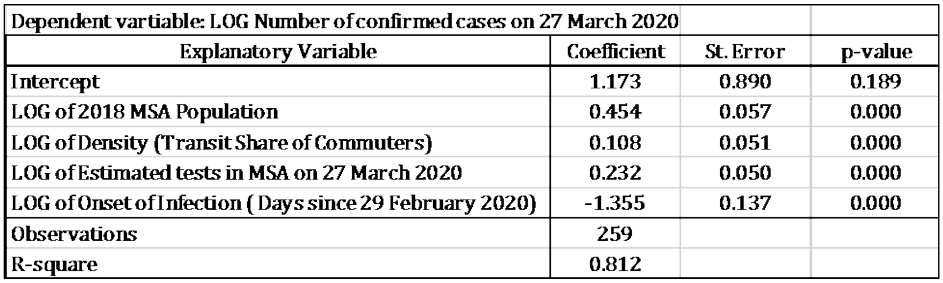
Explaining the Number of Reported Cases on 27 March 2020 in U.S. MSAs:

This section focuses on explaining the variation in the number of reported infections by the Coronavirus among U.S. Metropolitan Statistical Areas (MSAs). There is no doubt that the number of reported cases is lower than the real number of infected persons, and without extensive testing (or random testing of the MSA population) it is not possible to know that number.

We hypothesize that the number of reported cases would be higher in more populated metropolitan areas, in more dense metropolitan areas, in metropolitan areas with more extensive testing, and in metropolitan areas with an earlier onset of infection. We tested these hypotheses with the multiple regression model that appears in table 2 below. This model is surprisingly powerful: It explains 80 percent of the variation in the number of infections reported on 27 March 2020 in U.S. Metropolitan Statistical Areas (MSAs). All the explanatory variables in the model are significant at the 95% confidence level.

Table 2 is again displayed in logarithmic form. The logarithmic form of the model allows us to describe the coefficient of each of the explanatory variable as an elasticity: (1) a 10% increase in the total population of an MSA is associated with a 4.5% increase in the number of reported cases of infection (in this case by 27 March 2020); (2) a 10% increase in high-density living (measured by the share of commuters in the MSA using transit) is associated with a 1.1% increase in the number of reported cases of infection (in this case by 27 March 2020); (3) a 10% increase in the number of days since the onset of infection is associated with a 13.6% increase in the number of infections (in this case by 27 March 2020); and (4) a 10% increase in the number of tests for the Coronavirus conducted in the MSA (in this case by 27 March 2020) is associated with a 2.3% increase in the reported number of Coronavirus infections in the MSA.

The key finding here is that the reported number of Coronavirus infections in a given MSA is also predictable. The most powerful predictor is the number of days since the onset of infection in the MSA: The key reason that some MSAs have higher levels of reported infections is that the onset of reported infections in these MSAs occurred earlier. Second, other things being equal, larger and denser MSAs can be expected to have higher levels of reported infections. Third, MSAs with higher levels of testing are likely to identify more confirmed cases of infection.



Dependent vartiable: LOG Number of confirmed cases on 27 March 2020			
Explanatory Variable	Coefficient	St. Error	p-value
Intercept	1.173	0.890	0.189
LOG of 2018 MSA Population	0.454	0.057	0.000
LOG of Density (Transit Share of Commuters)	0.108	0.051	0.000
LOG of Estimated tests in MSA on 27 March 2020	0.232	0.050	0.000
LOG of Onset of Infection (Days since 29 February 2020)	-1.355	0.137	0.000
Observations	259		
R-square	0.812		

Table 2: Multiple regression model explaining the variation in the reported Coronavirus infections as of 27 March 2020 in U.S. MSAs.

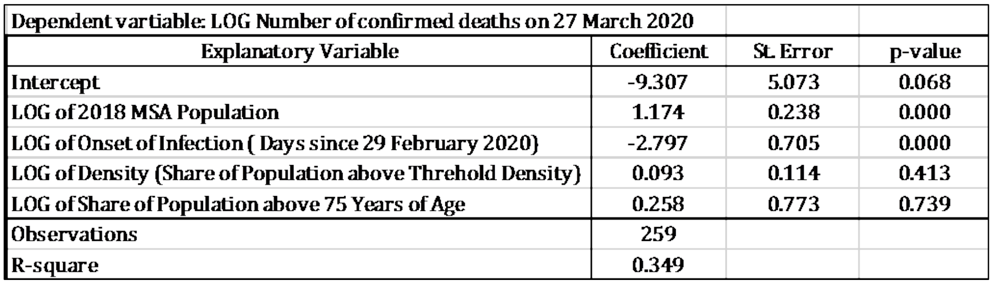
Explaining the Number of Deaths on 27 March 2020 in U.S. MSAs:

This section focuses on explaining the variation in the number of reported deaths from the Coronavirus infection among U.S. Metropolitan Statistical Areas (MSAs). The number of reported deaths is a more accurate measure than the number of reported infections, with the proviso that medical personnel can attribute the death to a Coronavirus infection.

We hypothesize that the number of reported deaths would be higher in more populated metropolitan areas, in more dense metropolitan areas, in metropolitan areas with a higher share of older people, in metropolitan areas with an earlier onset of infection and, of course, in metropolitan areas with higher levels of reported infections. We tested these hypotheses with two separate multiple regression model that appear in tables 3 and 4 below. The model in table 3 is not as powerful as the model in table 2 above, but it is still quite powerful: It explains 35 percent of the variation in the number of reported deaths as of 27 March 2020 in U.S. Metropolitan Statistical Areas (MSAs).

Table 3 is again displayed in logarithmic form. Only two of the explanatory variables in the model are significant at the 95% confidence level: The total population of the MSA and the date of the onset of reported infections. The logarithmic form of the model allows us to describe the coefficient of each of these two explanatory variable as an elasticity: (1) a 10% increase in the total population of an MSA is associated with an 12% increase in the number of reported deaths (in this case by 25 March 2020); (2) a 10% increase in the number of days since the onset of infection is associated with a 28.0% increase in the number of reported deaths (in this case by 27 March 2020). The two other variables in the model—the share of the population over 75 years of age and the share of the population living at high density (above 10,000 persons per square mile)—have the right effect but are not statistically significant. In other words, contrary to our expectations, the share of older people in an MSA population does not explain the number of deaths there.

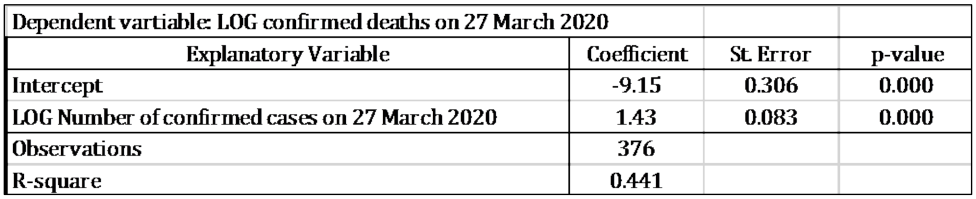
The key finding here is that the reported number of Coronavirus deaths in a given MSA is predictable. The most powerful predictor is the number of days since the onset of infection in the MSA: The key reason that some MSAs have higher levels of reported deaths is that the onset of reported infections in these MSAs occurred earlier. Second, other things being equal, MSAs with larger populations can expect to have higher levels of reported Coronavirus-related deaths.



Dependent vartiable: LOG Number of confirmed deaths on 27 March 2020			
Explanatory Variable	Coefficient	St. Error	p-value
Intercept	-9.307	5.073	0.068
LOG of 2018 MSA Population	1.174	0.238	0.000
LOG of Onset of Infection (Days since 29 February 2020)	-2.797	0.705	0.000
LOG of Density (Share of Population above Threhold Density)	0.093	0.114	0.413
LOG of Share of Population above 75 Years of Age	0.258	0.773	0.739
Observations	259		
R-square	0.349		

Table 3: Multiple regression model explaining the variation in Coronavirus-related deaths as of 27 March 2020 in U.S. MSAs.

The statistical model shown in table 3 did not include one important explanatory variable: The number of reported Coronavirus infections in a given MSA at a given date: The more people are infected the more people are likely to die. It is quite evident to everyone following the numbers reported daily in the media that the number of reported Coronavirus deaths is a relatively fixed share of the number of reported infections. Table 4 presents a regression model that seeks to explain the variation in the number of Coronavirus-related deaths at a given date simply as a function of the number of reported infections at that date.



Dependent vartiable: LOG confirmed deaths on 27 March 2020			
Explanatory Variable	Coefficient	St. Error	p-value
Intercept	-9.15	0.306	0.000
LOG Number of confirmed cases on 27 March 2020	1.43	0.083	0.000
Observations	376		
R-square	0.441		

Table 4: Simple regression model explaining the variation in Coronavirus-related deaths as a function of reported Coronavirus infections as of 27 March 2020 in U.S. MSAs.

Table 4 is again displayed in logarithmic form. The only one explanatory variable in the model—the confirmed number of infections as of 27 March 23020—is significant at the 95% confidence level, and it alone explains 44% in the variation in the number of deaths among

MSAs. The logarithmic form of the model allows us to describe its coefficient as an elasticity: A 10% increase in the number of reported Coronavirus infections (in this case by 27 March 2020) is associated with a 14.3% increase in the number of reported deaths (also by 27 March 2020).

Conclusion:

This is our first attempt to explain the variations in the date of the onset of infections, in the number of confirmed cases, and in the number of deaths from the Coronavirus in U.S. MSAs. Since the data is now reported daily, we plan to update and refine the model in the coming days.

The most important conclusion of our preliminary analysis is that the geographic spread of the Coronavirus in U.S. Metropolitan Statistical Areas (MSAs) is quite predictable and explainable. The variations in the date of the onset of infection in different MSAs can be adequately explained by their population size, their density, and their airline connections with other countries. The main reason that some MSAs report more infections than others is that the onset of infection there occurred earlier. In this sense, New York is not the *epicenter* of the pandemic but—together with Los Angeles, San Francisco, San Jose and Seattle—but our *vanguard,* our *avant-garde*, on the pandemic front. Secondary reasons that some MSAs report more infections than others are that they are larger and denser, and not necessarily because they contain a larger share of older people. All of this may be quite obvious, but in these times of uncertainty it may make provide some people some comfort to know that, for now, the onset of infection as well as the number of people infected and the number of people dying is explainable and, to an extent, even predictable. In subsequent analyses, we plan to update the models and possibly make them more comprehensive by including other factors in our analysis, such as per capita public health expenditures or the onset of state stay-at-home orders. We also plan to track historical data that allow us to focus on the rate of infection and the rate of dying—their change over time in each MSA—and try to explain variations in that rate among U.S. Metropolitan Statistical Areas.

* * *