Socioeconomic Factors influencing the Spatial Spread of COVID-19 in the United States*

Christopher F Baum (Boston College, DIW Berlin & CESIS)
Miguel Henry (Greylock McKinnon Associates)

2020-10-03

Abstract

As the COVID-19 pandemic has progressed in the U.S., "hotspots" have been shifting geographically over time to suburban and rural counties showing a high prevalence of the disease. We analyze daily U.S. county-level variations in COVID-19 confirmed case counts to evaluate the spatial dependence between neighboring counties. We find strong evidence of county-level socioeconomic factors influencing the spatial spread. We show the potential of combining spatial econometric techniques and socioeconomic factors in assessing the spatial effects of COVID-19 among neighboring counties.

Keywords: COVID-19, coronavirus, socioeconomic factors, spillover effects, spatial econometrics

JEL Classification: C13, C21, R15, R23

1 Introduction

"In a quickly changing pandemic landscape...county-level data and analysis is crucial to understanding needs and supporting planning efforts." [1]

In January 2020, the first confirmed case of COVID-19 in the U.S. was reported in the state of Washington (Holshue et al. 2020). The first U.S. deaths were officially reported in February in Washington by the CDC and in California.[2] By mid-March, COVID-19 transmission had become widespread, tending to cluster in certain subregions and accelerated with rapidly increasing case counts more than 1,000-fold with

*We thank participants in the International Workshop on Computational Economics and Econometrics, Rome, and the Stata Conference 2020, London for their comments.

[1] Center for Spatial Data Science, University of Chicago.

[2] See https://time.com/5825320/california-coronavirus-february-first-death/.

New York City being the outbreak's focal point, with other urban areas in the Northeast and Midwest seriously affected (see Appendix). Since then all 50 states and D.C. implemented various actions such as non-pharmaceutical public health interventions (PHIs) to slow down and contain the transmission of the ongoing coronavirus pandemic. These measures include social distancing, business closures, school closings, public gathering restrictions, travel limitations and stay-at-home orders.[3] Some local governments such as Sonoma County, CA have also implemented stay-at-home and shelter-in-place orders.[4]

A rapidly growing literature on spatial geographic aspects related to COVID-19 includes Bailey et al. (2020), Coven & Gupta (2020), Guliyev (2020), Kang et al. (2020), Kuchler et al. (2020), and Mollalo et al. (2020).

In the following, we briefly describe the data that we use in this study.

2 Data

We constructed a cross-sectional data set for counties of the 48 contiguous U.S. states and the District of Columbia for Spring 2020. Daily data on COVID-19 confirmed cases for each U.S. state and county were obtained from USAFacts.[5] The data include the state, the county, its Federal Information Processing System (FIPS) code, and the daily confirmed case counts of COVID-19 from March 1, 2020 until the present. Descriptive statistics on confirmed cases for county j and date t are presented in Table 1. The Appendix contains county-level choropleth maps showing the spread of COVID-19 cases for March 1, April 1, April 30 and the last date referenced in the analysis, May 23, 2020.

We obtain additional data from a variety of sources. Because of differences in population across counties, the number of confirmed COVID-19 cases is adjusted in each county by dividing the confirmed counts by the total population of each county in 2018, the latest county resident population estimates with demographic characteristics available from the U.S. Census Bureau. Using this same data source, we gathered information on gender and race at the county level across five aggregated age groups.[6] These data were included in the spatial models as covariates in percentage units. In addition to this demographic information, we added county-level data for 2018 on three socioeconomic factors: median income, prevalence of PM2.5 pollution[7] and the percent of residents lacking health insurance from the County Health Rankings database.

[3] A complete list of the interventions implemented in each state and their effective dates are available at the Institute for Health Metrics and Evaluation, University of Washington.

[4] See https://socoemergency.org/order-of-the-health-officer-shelter-in-place/.

[5] These data were originally obtained from Johns Hopkins Center for Systems Science and Engineering (CSSE). However, those data did not distinguish cases in the boroughs of New York City.

[6] Below 20 years, between 20 and 39 years, between 40 and 59 years, between 60 and 79 years, and 80 years or more.

[7] PM2.5 describes fine inhalable particles, with diameters 2.5 micrometers and smaller where PM denotes particulate matter.

Finally, we added county-level data from PolicyMap and CDC BRFSS (2020) for their COVID-19 Health Risk Index. This index incorporates the prevalence of five health conditions which have been considered as risk factors for COVID-19 infections[8] from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) at the county level, and is available in z-score form.

Table 2 provides descriptive statistics of the demographic, socioeconomic and health index risk variables that are used in the spatial econometric models presented below. The variables White, Black and Hispanic include both males and females but do not add up to 100%, as Hispanic ethnicity may be combined with any or several of the U.S. Census Bureau racial categories (White, Black or African American, American Indian or Alaska Native, Asian, and Native Hawaiian or other Pacific Islander).

Spatial data, including geographic coordinates and the FIPS identifier for each county for 2018, were obtained from the Census Bureau's MAF/TIGER geographic database.

3 Empirical Strategy and Results

In this section, we explore the potential of combining spatial econometric techniques and socioeconomic factors to assess the spatial effects of COVID-19 among neighboring counties.

To accommodate spatial autoregression in our models and measure spatial spillover effects of COVID-19, the $(N \times N)$ spatial matrix $\mathbf{W}$ was computed using rook contiguity for the 48 contiguous U.S. states and the District of Columbia, where N is the number of U.S. counties (FIPS) under study. This weighting matrix measure implies that counties are considered first-order neighbors if they share a border, and not merely a vertex. We also tested an inverse distance spatial matrix, which yielded less satisfactory results. Therefore, results reported in Tables 4 and 5 rely on the rook contiguity matrix rather than on the inverse distance matrix.

3.1 The spatial autoregressive (SAR) model

In econometrics, several estimation procedures have been developed to model spatial dependence and examine the spatial relationships among neighboring units. See Anselin (1988) for a comprehensive discussion of the use of various estimators (least squares, maximum likelihood, instrumental variable, and method of moments) to account for spatial correlation issues in the context of the linear regression model. LeSage & Pace (2009) provide a textbook introduction to the SAR model. Kelejian & Prucha (1998, 1999) examined the generalized two stage least squares and the generalized method of moments estimators, while Lee (2004) derived the properties of the maximum likelihood estimator and its robust covariance matrix. In this paper, we estimate

[8]Obesity, diabetes, high blood pressure, heart disease and chronic obstructive pulmonary disease. Asthma is not included in the health risk index due to data inconsistency on asthma risk.

SAR lag models using the maximum likelihood estimator provided by Stata version 16 as the `spregress` command.

We first estimated a SAR model in which the most recent confirmed case count $C_{j,t}$ for county j is modeled by $C_{j,t-14}$ and $C_{j,t-28}$, t = May 23. Table 1 summarizes these variables. These lag lengths were motivated by the incubation period of COVID-19 and the required isolation period of a fortnight. Considering that spatial dependencies occur through multiple channels, this model makes use of both a spatial lag of the dependent variable ($C_{j,t}$, confirmed case counts), specified by the spatial contiguity matrix **W**, as well as a spatially lagged error term utilizing the same spatial matrix **W**.

3.2 Incorporating socioeconomic factors

As a second step, we augmented the previous pure SAR model with several socioeconomic factors described in Table 2, motivated by the following stylized facts:

1. Gender: rejection of PHIs by males has been commonly observed in many settings, including lockdown protests in a number of states.
2. Race/ethnicity: minority workers have been more likely to be on the 'front lines' in many essential industries, and less likely to be able to work from home confinement.
3. Age: while COVID-19 infections are more serious for the aged, who are more likely to avoid infection if they can do so, younger individuals are more likely to ignore PHIs and contribute to the spread.
4. Income: counties with lower median income are likely to have lower-quality health care resources and a higher percentage of minority residents.
5. Pollution: residents of counties with higher levels of air pollution are likely to be more susceptible to airborne infection due to their exposure to pollutants.
6. Health insurance: residents of counties with a larger fraction lacking health insurance are likely to be more susceptible to COVID-19.
7. Complicating health conditions: residents of counties with higher health risk indices are likely to be more susceptible to COVID-19.

These are the rationales for the inclusion of the county-level fraction of males, the fractions of Blacks and Hispanics, the fraction in the 20–39 year age group,[9] the level of PM2.5, median income, the fraction of uninsured and the standardized health risk index in the estimated models. In addition, Persico & Johnson (2020) found a strong, positive relationship between pollution and COVID-19 mortality and case rates, especially in counties with higher populations of Black, lower income and unemployed

[9]Preliminary empirical investigations with the five age groups suggested that the only one that signals a clear high prevalence of cases is the 20–39 year age group, which explains its inclusion in the models.

individuals. With regard to the COVID-19 Health Risk index inclusion, descriptive evidence shows that chronic health conditions can exacerbate susceptibility to COVID-19.[10]

To illustrate the relationships among the socioecoonmic factors at the county level, Table 3 presents simple correlations among the factors included in our models.

3.3 Empirical results

In the SAR model presented in the first column of Table 4, where the confirmed case count in county j is affected by both the confirmed case counts and by the unobserved factors in county j's neighbors, the relevance of spatial factors over and above the autoregressive factors is evident. The model pv refers to the test of the overall model, whereas the spatial pv refers to a test of the relevance of spatial factors.

The models incorporating socioeconomic factors in Tables 4 and 5 include only one autoregressive term, $C_{j,t-14}$. That factor is statistically significant at the 99% level in all estimated models. Its value, in excess of unity, reflects the exponential growth potential of the pandemic, which even in the presence of preventive public health measures is an explosive dynamic process in the absence of a reservoir of uninfected individuals.

In column 2 of Table 4, gender composition for each county is added. Counties with higher percentages of male residents are predicted to have significantly higher confirmed cases, *ceteris paribus*. Column 3 includes the two race and ethnicity measures, reflecting the minority composition of the county. They each increase predicted case counts, and are jointly statistically significant. Column 4 illustrates the positive impact of a larger number of residents aged 20–39. Column 5 combines the gender and minority status variables, further increasing the log-likelihood statistic. Finally, column 6 reflects the impact of county median income level on confirmed case counts. As expected, higher median income is associated with lower confirmed cases, *ceteris paribus*. These models thus provide support for the first four stylized facts listed above. Over and above the SAR effects, which are uniformly significant, these socioeconomic factors have strong effects on the prevalence of COVID-19 infections.

In column 1 of Table 5, the impact of air pollution levels, proxied by PM2.5, significantly increases the confirmed case count. The second column reproduces the effect of median income for comparison. In column 3, the impact of residents lacking health insurance coverage is positive and statistically significant. Column 4 include the Health Risk Index measure, where higher levels of the index reflect greater prevalence of five health conditions. The impact on confirmed cases is positive and significant. Column 4 combines the air pollution measure with the percentage of residents lacking health insurance, with those factors individually and jointly significant. Finally, column 6 combines the air pollution measure, which is often higher in minority neighborhoods, with the race and ethnicity factors. Each of these variables has a significant

[10] See https://www.nytimes.com/interactive/2020/05/18/us/coronavirus-underlying-conditions.html.

effect on $C_{j,t}$, individually and in combination. Models incorporating these socioeconomic factors have considerably lower log-likelihood values than the pure spatial autoregressive model presented in Table 4. When added to the SAR framework, the socioeconomic factors have a clear impact on the prevalence of COVID-19 infections, supporting stylized facts 5–7 listed above.

The statistics shown under the `CDIST` heading, `Ct` and `e.Ct`, show the impact of the spatial lag of the dependent variable $C_{j,t}$ and the spatial error lag, respectively. The former lacks statistical significance in some of the models, but it is significantly different from zero in most SAR models and it has a positive sign across all models. This indicates that neighboring counties have a significant, positive effect on the COVID-19 confirmed counts, implying that we cannot ignore the spatial impacts from neighbors.[11] The spatial error lag coefficients are always significantly different from zero. The spatial pv is the p-value of a χ^2 test for the relevance of spatial factors, which strongly rejects its null hypothesis for all models considered in this paper.

4 Conclusions and Extensions

This first foray into spatial autoregressive modeling of the COVID-19 pandemic in the United States reveals the usefulness of this modeling framework, capturing both geography and socioeconomic factors as important contributors to the severity of the pandemic. As a proof of concept, it illustrates the potential for geographic modeling to enhance our understanding of a fast-moving dynamic process. Socioeconomic factors, including demographics and health risk measures, change at a much slower rate. Nevertheless, these quasi-fixed factors can add significantly to our understanding of the spread of COVID-19.

Looking ahead, we plan to extend the current econometric cross-sectional framework to spatial autoregressive panel data models with fixed effects along both spatial and sociodemographic lines to further investigate their potential. We also hope to complement the econometric analysis with a local spatial autocorrelation analysis based on the Getis & Ord (1992) hot and cold spot approach and on population-weighted coordinates in addition to the Moran (1950) *I* global statistic and geographic coordinates.

[11] Although not reported in this paper, we estimated the Moran (1950) *I* statistic to formally test for global spatial clustering or geographic connectivity among the counties under study. The resulting p-value of zero to five decimal places strongly rejected the null hypothesis of spatial randomness.

References

Anselin, L. (1988), *Spatial Econometrics: Methods and Models*, Kluwer Academic Publishers, Dordrecht.

Bailey, M., Kuchler, T., Russel, D., State, B. & Stroebel, J. (2020), Social Connectedness in Europe, Working paper, Stern School of Business, New York University.

Coven, J. & Gupta, A. (2020), Disparities in Mobility Responses to COVID-19, Working paper, Stern School of Business, New York University.

Getis, A. & Ord, J. K. (1992), 'The analysis of spatial association by use of distance statistics', *Geographical Analysis* **24**, 189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x.

Guliyev, H. (2020), 'Determining the spatial effects of COVID-19 using the spatial panel data model', *Spatial Statistics* **38**, 1–10. https://www.sciencedirect.com/science/article/pii/S2211675320300373.

Holshue, M., DeBolt, C., Lindquist, S. & et al. (2020), 'First Case of 2019 Novel Coronavirus in the United States', *New England Journal of Medicine* **382**, 929–936. http://hdl.handle.net/10.1056/NEJMoa2001191.

Kang, D., Choi, H., Kim, J.-H. & Choi, J. (2020), 'Spatial epidemic dynamics of the COVID-19 outbreak in China', *International Journal of Infectious Diseases* **94**, 96–102. https://www.sciencedirect.com/science/article/pii/S1201971220302095.

Kelejian, H. & Prucha, I. (1998), 'A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances', *Journal of Real Estate Finance and Economics* **17**, 99–121. https://link.springer.com/article/10.1023/A:1007707430416.

Kelejian, H. & Prucha, I. (1999), 'A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model', *International Economic Review* **40**, 509–533. https://www.jstor.org/stable/2648817.

Kuchler, T., Russel, D. & Stroebel, J. (2020), The Geographic Spread of COVID-19 Correlates with Structure of Social Networks as Measured by Facebook, Cesifo working paper no. 8241, CESifo. https://ideas.repec.org/p/ces/ceswps/_8241.html.

Lee, L.-F. (2004), 'Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models', *Econometrica* **72**, 1899–1925. http://hdl.handle.net/10.1111/j.1468-0262.2004.00558.x.

LeSage, J. & Pace, R. K. (2009), *Introduction to Spatial Econometrics*, Boca Raton, FL: Chapman & Hall/CRC.

Mollalo, A., Vahedi, B. & Rivera, K. (2020), 'GIS-based spatial modeling of COVID-19 incidence rate in the continental United States', *Science of the Total Environment* **728**, 138884. https://doi.org/10.1016/j.scitotenv.2020.138884.

Moran, P. (1950), 'Notes on Continuous Stochastic Phenomena', *Biometrika* **37**, 17–23. https://pdfs.semanticscholar.org/56a3/ebf5aa12dc98f13ae34d25c0eb0ed4ae4f32.pdf.

Persico, C. & Johnson, K. (2020), Deregulation in a Time of Pandemic: Does Pollution Increase Coronavirus Cases or Deaths?, IZA DP No. 13231, IZA. https://ideas.repec.org/p/iza/izadps/dp13231.html.

PolicyMap and CDC BRFSS (2020), COVID Risk Index, Technical report, PolicyMap. https://PolicyMap.com.

Table 1: Confirmed COVID-19 Counts for Lag Variables (N = 3,107)

Variable Name	Min	0.25	Median	0.75	0.99	Max	Date
$C_{j,t}$	0	7	33	157	8,316	197,266	May 23
$C_{j,t-14}$	0	5	23	107	6,887	183,289	May 9
$C_{j,t-28}$	0	3	14	68	4,916	155,113	April 25

Table 2: Descriptive Statistics (N = 3,107)

Variable Name	Mean	Min	Max	Description
pct_male	50.08	43.13	73.16	Percentage of Males
pct_black	10.30	0.097	86.61	Percentage of Blacks
pct_hisp	9.69	0.61	96.36	Percentage of Hispanics
pct_20_30	24.01	11.30	53.06	Percentage in the 20–39 year age group
pm25	8.95	4.2	15.4	PM2.5 (in μg per cubic meter)
medinc	49.39	22.05	134.61	Median Income ('000 US dollars)
unins	11.95	2.13	33.27	Percentage of Uninsured Individuals
index_zscore	0.08	-3.65	3.56	z–score of Health Risk Index value

Notes: The sample size N denotes the number of U.S. counties (FIPS) under study.

Table 3: Correlations of regressors

	pct male	pct black	pct hisp	pct 20 30	pm25	medinc	unins	index zscore
pct male	1.000							
pct black	-0.145	1.000						
pct hisp	0.160	-0.093	1.000					
pct 20 30	0.213	0.278	0.233	1.000				
pm25	-0.253	0.247	-0.265	0.195	1.000			
medinc	-0.036	-0.237	0.040	0.116	0.070	1.000		
unins	0.065	0.191	0.451	0.006	-0.275	-0.395	1.000	
index zscore	-0.129	0.352	-0.291	-0.330	0.256	-0.595	0.191	1.000

Table 4: Spatial models of COVID-19 Cases: I

	(1)	(2)	(3)	(4)	(5)	(6)
Ct						
Ct14	1.182***	1.137***	1.129***	1.136***	1.125***	1.141***
	(124.33)	(179.38)	(177.77)	(177.89)	(177.07)	(180.39)
Ct28	-0.102***					
	(-6.01)					
pct_male		5.982***			6.275***	
		(4.75)			(5.00)	
pct_black			1.926***		2.040***	
			(8.07)		(8.51)	
pct_hisp			1.385***		1.258***	
			(5.82)		(5.25)	
pct_20_30				3.303***		
				(4.84)		
medinc						-0.900***
						(-3.67)
CDIST						
Ct	0.0483***	0.0305***	0.00480	0.0237**	0.00938	0.0318***
	(4.06)	(2.58)	(0.40)	(2.02)	(0.78)	(2.70)
e.Ct	0.213***	0.275***	0.287***	0.270***	0.293***	0.262***
	(6.34)	(8.30)	(8.63)	(8.13)	(8.85)	(7.88)
N	3107	3107	3107	3107	3107	3107
LogLikelihood	-19999.7	-20005.4	-19971.2	-20005.0	-19958.8	-20009.9
PseudoR2	0.924	0.923	0.925	0.923	0.926	0.923
model pv	0	0	0	0	0	0
spatial pv	4.20e-18	7.03e-22	4.54e-19	8.73e-20	1.30e-20	2.55e-20

t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Spatial models of COVID-19 Cases: II

	(1)	(2)	(3)	(4)	(5)	(6)
Ct						
Ct14	1.139***	1.141***	1.139***	1.141***	1.137***	1.128***
	(179.83)	(180.39)	(180.36)	(179.93)	(180.16)	(177.48)
pm25	5.448**				8.649***	5.493**
	(2.46)				(3.75)	(2.37)
medinc		-0.900***				
		(-3.67)				
unins			2.935***		3.618***	
			(4.45)		(5.29)	
index_zscore				8.449***		
				(2.69)		
pct_black						1.831***
						(7.56)
pct_hisp						1.511***
						(6.20)
CDIST						
Ct	0.0204*	0.0318***	0.0267**	0.0255**	0.0195	0.00132
	(1.71)	(2.70)	(2.27)	(2.19)	(1.63)	(0.11)
e.Ct	0.269***	0.262***	0.273***	0.257***	0.279***	0.290***
	(8.09)	(7.88)	(8.23)	(7.73)	(8.39)	(8.71)
N	3107	3107	3107	3107	3107	3107
LogLikelihood	-20013.6	-20009.9	-20006.8	-20013.0	-19999.7	-19968.4
PseudoR2	0.923	0.923	0.923	0.923	0.924	0.925
model pv	0	0	0	0	0	0
spatial pv	3.47e-19	2.55e-20	8.90e-21	1.60e-18	2.57e-20	4.32e-19

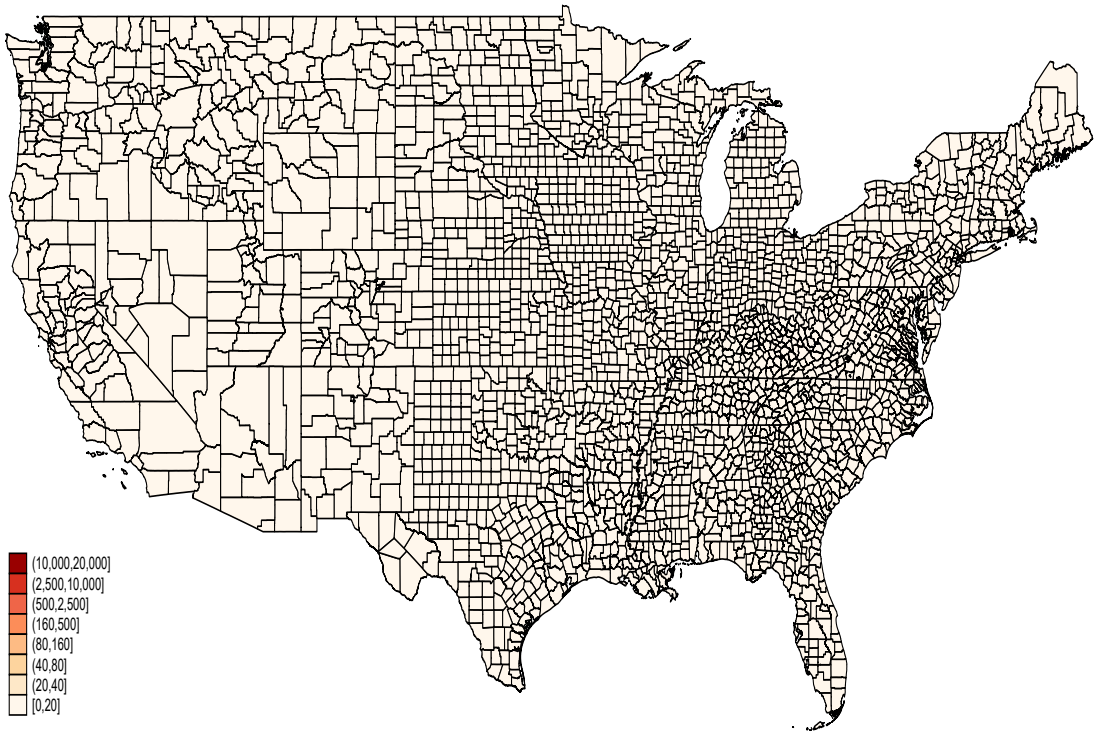
t statistics in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Appendix

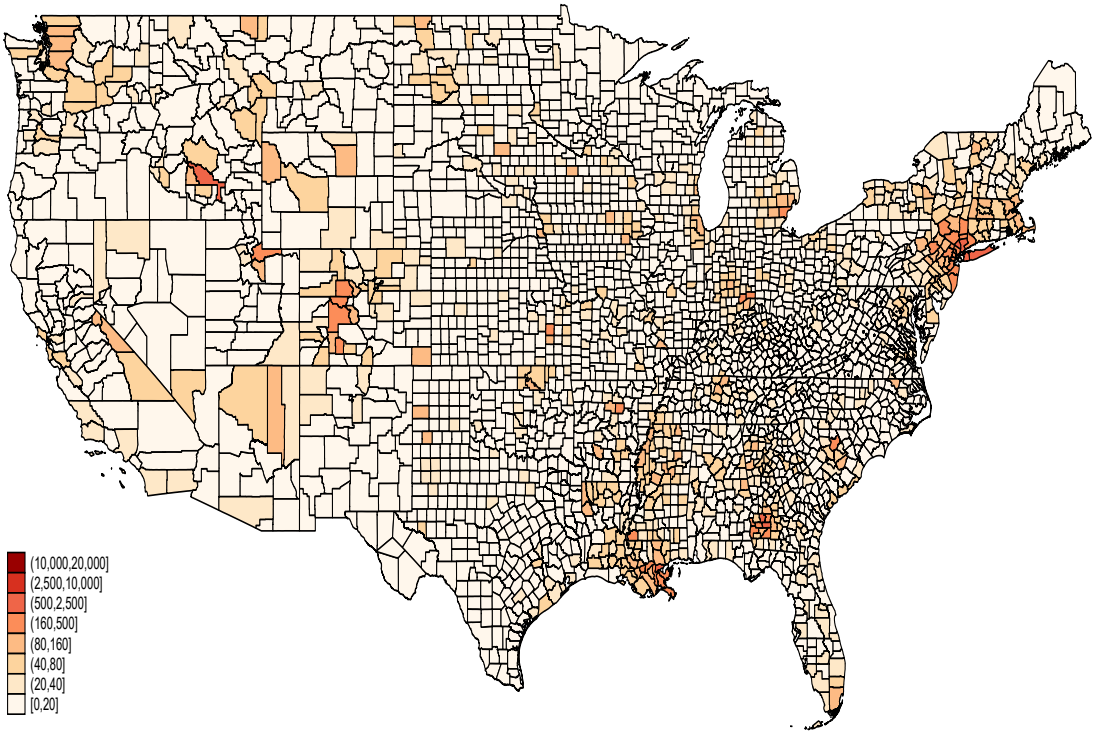
Confirmed Cases of COVID-19 in the United States

(cases per 100,000 population as of 3/1/2020)



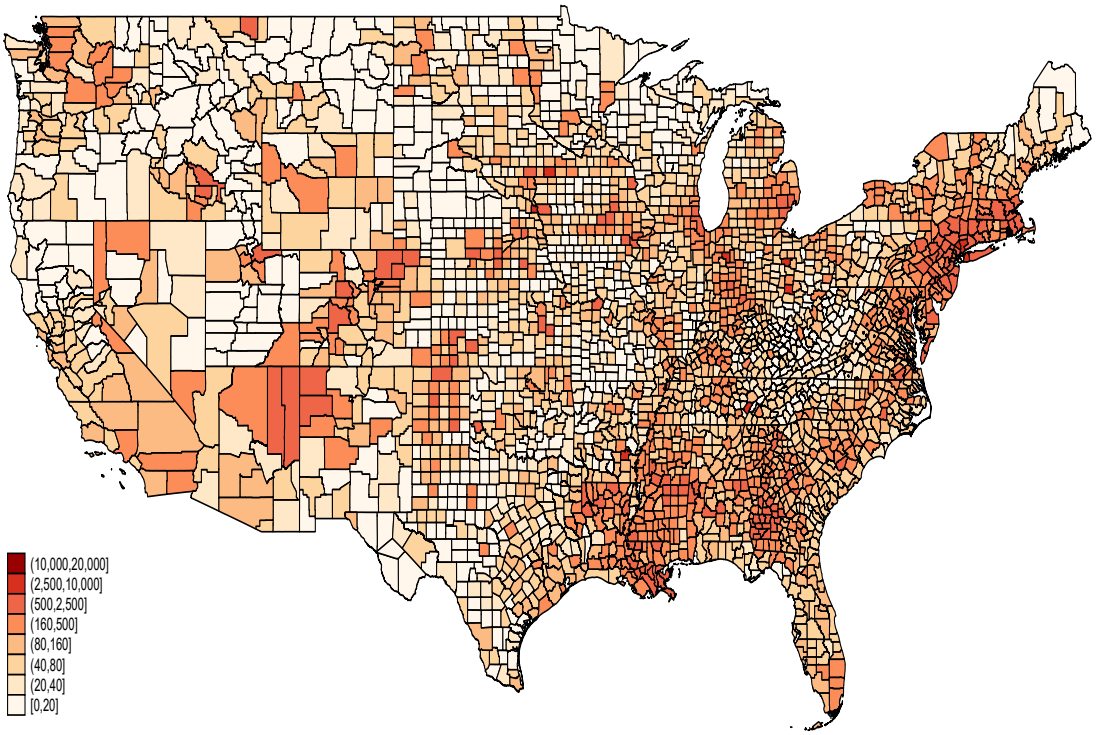
Confirmed Cases of COVID-19 in the United States

(cases per 100,000 population as of 4/1/2020)



Confirmed Cases of COVID-19 in the United States

(cases per 100,000 population as of 4/30/2020)



Confirmed Cases of COVID-19 in the United States

cases per 100,000 population as of 5/23/2020