COVID-19 spatial panel autoregressive modeling with US Household Pulse survey data

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

London Stata Conference, September 2021

Motivation for the study

- The spread of COVID-19 in the US has been heavily influenced by geography and proximity to areas with high concentrations of those infected
- A number of socioeconomic factors and state-level policies have emerged as predictors of the pandemic's severity
- These aspects warrant further investigation of the transmission process using appropriate econometric methods
- This paper analyzes 14 months of the pandemic's spread using the US Census Bureau's Household Pulse survey data
- Panel spatial autoregression techniques allow for time variation in confirmed case rates and death rates at the state level

Motivation for the study

- The spread of COVID-19 in the US has been heavily influenced by geography and proximity to areas with high concentrations of those infected
- A number of socioeconomic factors and state-level policies have emerged as predictors of the pandemic's severity
- These aspects warrant further investigation of the transmission process using appropriate econometric methods
- This paper analyzes 14 months of the pandemic's spread using the US Census Bureau's Household Pulse survey data
- Panel spatial autoregression techniques allow for time variation in confirmed case rates and death rates at the state level

Motivation for the study

- The spread of COVID-19 in the US has been heavily influenced by geography and proximity to areas with high concentrations of those infected
- A number of socioeconomic factors and state-level policies have emerged as predictors of the pandemic's severity
- These aspects warrant further investigation of the transmission process using appropriate econometric methods
- This paper analyzes 14 months of the pandemic's spread using the US Census Bureau's Household Pulse survey data
- Panel spatial autoregression techniques allow for time variation in confirmed case rates and death rates at the state level

Motivation for the study

- The spread of COVID-19 in the US has been heavily influenced by geography and proximity to areas with high concentrations of those infected
- A number of socioeconomic factors and state-level policies have emerged as predictors of the pandemic's severity
- These aspects warrant further investigation of the transmission process using appropriate econometric methods
- This paper analyzes 14 months of the pandemic's spread using the US Census Bureau's Household Pulse survey data
- Panel spatial autoregression techniques allow for time variation in confirmed case rates and death rates at the state level

Motivation for the study

- The spread of COVID-19 in the US has been heavily influenced by geography and proximity to areas with high concentrations of those infected
- A number of socioeconomic factors and state-level policies have emerged as predictors of the pandemic's severity
- These aspects warrant further investigation of the transmission process using appropriate econometric methods
- This paper analyzes 14 months of the pandemic's spread using the US Census Bureau's Household Pulse survey data
- Panel spatial autoregression techniques allow for time variation in confirmed case rates and death rates at the state level

Introduction

This study presents a preliminary investigation of the spread of COVID-19 in the US from late April, 2020 through early July, 2021 using spatial autoregression techniques.

Unlike our earlier study based on daily data at the US county level presented at IWEBcee 2020 (Baum and Henry, *IJCEE* forthcoming), we analyze state-level data aligned with 'waves' of the US Census Bureau's Household Pulse survey of more than 2.5 million respondents. This forms a panel of 49 states (including DC, excluding Alaska and Hawaii) for 27 two-week time periods.

Introduction

This study presents a preliminary investigation of the spread of COVID-19 in the US from late April, 2020 through early July, 2021 using spatial autoregression techniques.

Unlike our earlier study based on daily data at the US county level presented at IWEBcee 2020 (Baum and Henry, *IJCEE* forthcoming), we analyze state-level data aligned with 'waves' of the US Census Bureau's Household Pulse survey of more than 2.5 million respondents. This forms a panel of 49 states (including DC, excluding Alaska and Hawaii) for 27 two-week time periods.

We originally planned to employ heterogeneous spatial regression techniques which allow for the strength of spatial relationships to vary across panel units, following Aquaro, Bailey and Pesaran (*J.Applied Econometrics*, 2021) and their implementation of a quasi-maximum likelihood estimator.

Although there are preliminary implementations of the Aquaro et al. models for several programming languages, we encountered computational problems in both the Stata and Python versions of these packages. Those implementations also lack the equivalent of the `estat impact` postestimation routine, which computes the sum of direct and indirect effects on the outcome.

Consequently, we estimated spatial panel regression models at the national level and separately for each of the four US Census regions, allowing for heterogeneity across groups of contiguous states.

We originally planned to employ heterogeneous spatial regression techniques which allow for the strength of spatial relationships to vary across panel units, following Aquaro, Bailey and Pesaran (*J.Applied Econometrics*, 2021) and their implementation of a quasi-maximum likelihood estimator.

Although there are preliminary implementations of the Aquaro et al. models for several programming languages, we encountered computational problems in both the Stata and Python versions of these packages. Those implementations also lack the equivalent of the `estat impact` postestimation routine, which computes the sum of direct and indirect effects on the outcome.

Consequently, we estimated spatial panel regression models at the national level and separately for each of the four US Census regions, allowing for heterogeneity across groups of contiguous states.

We originally planned to employ heterogeneous spatial regression techniques which allow for the strength of spatial relationships to vary across panel units, following Aquaro, Bailey and Pesaran (*J.Applied Econometrics*, 2021) and their implementation of a quasi-maximum likelihood estimator.

Although there are preliminary implementations of the Aquaro et al. models for several programming languages, we encountered computational problems in both the Stata and Python versions of these packages. Those implementations also lack the equivalent of the `estat impact` postestimation routine, which computes the sum of direct and indirect effects on the outcome.

Consequently, we estimated spatial panel regression models at the national level and separately for each of the four US Census regions, allowing for heterogeneity across groups of contiguous states.

Spatial panel regression

In the econometric literature, several estimation procedures have been developed to model the spatial relationships among neighboring units. See Anselin (*Spatial Econometrics*, 1988, and chapter in *Companion to Theoretical Econometrics*, ed. Baltagi, 2003) for a comprehensive discussion of the use of various estimation techniques (least squares, maximum likelihood, instrumental variable, and method of moments) to account for spatial autocorrelation (SAR, spatial dependence) or structural instability (spatial heterogeneity) issues in the context of the linear regression model.

SAR models have attracted significant attention in this field since Cliff and Ord (*Spatial Processes*, 1981) due to their parsimonious representation of the cross-sectional correlation by a spatial weighting matrix (Bao, Liu, Yang, *Econometrics*, 2020) that plays an important role in describing the structure of the underlying spatial autoregressive data generating process. LeSage and Pace (*Introduction to Spatial Econometrics*, 2009) provide a textbook introduction to the SAR model.

To accommodate spatial dependence in our econometric models and measure spatial spillover effects, the $(N \times N)$ spatial weight matrix $\mathbf{W}$ was computed using queen contiguity for the 48 contiguous U.S. states and the District of Columbia, where N=49. This spatial weighting matrix measure implies that states are considered first-order neighbors if they share a vertex.

To guarantee nonsingularity and estimability of $\mathbf{W}$ as well as interpretability of the spatial lag model coefficients, $\mathbf{W}$ was normalized using spectral normalization so that its largest eigenvalue is 1. See Kelejian and Prucha (*J.Econometrics*, 2010) for an introduction to the use and interpretation of normalization methods on weighting matrices.

To accommodate spatial dependence in our econometric models and measure spatial spillover effects, the $(N \times N)$ spatial weight matrix $\mathbf{W}$ was computed using queen contiguity for the 48 contiguous U.S. states and the District of Columbia, where N=49. This spatial weighting matrix measure implies that states are considered first-order neighbors if they share a vertex.

To guarantee nonsingularity and estimability of $\mathbf{W}$ as well as interpretability of the spatial lag model coefficients, $\mathbf{W}$ was normalized using spectral normalization so that its largest eigenvalue is 1. See Kelejian and Prucha (*J.Econometrics*, 2010) for an introduction to the use and interpretation of normalization methods on weighting matrices.

We estimate models for the COVID-19 confirmed case rate and death rate from state-level panel data using the random-effects estimator available in Stata version 17: `spxtregress, re`. This extension of the linear random-effects panel model is estimated by maximum likelihood as described by Lee and Yu (*Regional Science and Urban Economics*, 2010) based on the earlier work of Kapoor, Kelejian and Prucha (*J.Econometrics*, 2007).

In the random effects spatial autoregressive model for panel data, the random effects enter the equation for y_{nt} linearly.

$$\begin{aligned} y_{nt} &= \lambda \mathrm{W} y_{nt} + \mathrm{Z}_{nt}\beta + c_n + u_{nt} \\ u_{nt} &= \rho \mathrm{W} u_{nt} + v_{nt}, \quad t = 1, 2, \ldots, T \end{aligned}$$

where Z_{nt} may contain time-varying and time-invariant regressors, possibly including their spatial lags,
c_n are random effects distributed $(0, \sigma_c^2)$,
u_{nt} is a $n \times 1$ vector of spatially lagged errors,
v_{nt} is a $n \times 1$ vector of innovations, $i.i.d.$ across i and t with variance σ^2,
W is the spatial weighting matrix.

Data for the T time periods can be stacked to write the equations as a $nT \times 1$ vector

$$y_{nT} = \lambda(\mathrm{I} \otimes \mathrm{W})y_{nT} + \mathrm{Z}_{nT}\beta + \zeta_{nT}$$

where the overall disturbance vector ζ_{nT} is

$$\begin{aligned} \zeta_{nT} &= \mathrm{I}_T \otimes c_n + (\mathrm{I}_T \otimes \mathrm{R}_n(\rho)^{-1})v_{nT} \\ \mathrm{R}_n(\rho) &= \mathrm{I}_n - \rho\mathrm{W} \end{aligned}$$

The variance matrix of the estimator is then

$$\Omega_{nT}(\theta) = \sigma_c^2(\iota_T\iota_T' \otimes \mathrm{I}_T) + (\mathrm{I}_T \otimes \mathrm{R}_n(\rho)^{-1} \otimes \mathrm{R}_n'(\rho)^{-1})$$

where ι_T is a $T \times 1$ vector of 1s and $\theta = (\beta', \lambda, \rho, \sigma_c^2, \sigma^2)'$. The Ω matrix can then be used, assuming $i.i.d.$ disturbances, to define the log-likelihood function to be maximized.

Data for the T time periods can be stacked to write the equations as a $nT \times 1$ vector

$$y_{nT} = \lambda(\mathsf{I} \otimes \mathsf{W})y_{nT} + \mathsf{Z}_{nT}\beta + \zeta_{nT}$$

where the overall disturbance vector ζ_{nT} is

$$\begin{aligned} \zeta_{nT} &= \mathsf{I}_T \otimes c_n + (\mathsf{I}_T \otimes \mathsf{R}_n(\rho)^{-1})v_{nT} \\ \mathsf{R}_n(\rho) &= \mathsf{I}_n - \rho\mathsf{W} \end{aligned}$$

The variance matrix of the estimator is then

$$\Omega_{nT}(\theta) = \sigma_c^2(\iota_T\iota_T' \otimes \mathsf{I}_T) + (\mathsf{I}_T \otimes \mathsf{R}_n(\rho)^{-1} \otimes \mathsf{R}_n'(\rho)^{-1})$$

where ι_T is a $T \times 1$ vector of 1s and $\theta = (\beta', \lambda, \rho, \sigma_c^2, \sigma^2)'$. The Ω matrix can then be used, assuming $i.i.d.$ disturbances, to define the log-likelihood function to be maximized.

The parameter ρ gauges the strength of spatial dependence in the dependent variable. The coefficients in β do not provide the effects of a change in an explanatory variable Z on y, as those changes vary over all spatial units. Both the β and ρ coefficients must be used to compute the direct and indirect impact effects (spatial spillovers) of a change in Z, taking the recursive effects into account. The reduced form conditional mean is then given by

$$\mathrm{E}[\mathrm{y}|\mathrm{Z},\mathrm{W}] = (\mathrm{I} - \rho\mathrm{W})^{-1}\mathrm{Z}\beta \tag{1}$$

where I is an *N*-dimensional identity matrix and E[.] is the conditional mean.

The direct effects capture the contributions of each unit's values of Z on its reduced form mean, while the indirect effects capture the contributions of the other units' values of Z on each unit's reduced form mean.

In essence, the average direct and indirect effects notion corresponds to average partial derivatives, and the sum of these two is the average total effect, or average total impact, of the Z variables on y. These are calculated with Stata's `estat impact` command.

In the tables below, we present estimates of the average total impact, with standard errors computed via the delta method.

The direct effects capture the contributions of each unit's values of Z on its reduced form mean, while the indirect effects capture the contributions of the other units' values of Z on each unit's reduced form mean.

In essence, the average direct and indirect effects notion corresponds to average partial derivatives, and the sum of these two is the average total effect, or average total impact, of the Z variables on y. These are calculated with Stata's `estat impact` command.

In the tables below, we present estimates of the average total impact, with standard errors computed via the delta method.

The direct effects capture the contributions of each unit's values of Z on its reduced form mean, while the indirect effects capture the contributions of the other units' values of Z on each unit's reduced form mean.

In essence, the average direct and indirect effects notion corresponds to average partial derivatives, and the sum of these two is the average total effect, or average total impact, of the Z variables on y. These are calculated with Stata's `estat impact` command.

In the tables below, we present estimates of the average total impact, with standard errors computed via the delta method.

For brevity, we do not present the estimated direct and indirect effects, but in every case where the total impact is significant at the 95% level of confidence, the indirect effects are themselves significant.

The 'spatial pv' is the p-value of a Wald test of the null hypothesis that spatial effects do not contribute to the explanatory power of the model. The 'model pv' is the standard Wald statistic for the overall explanatory power of the model.

For brevity, we do not present the estimated direct and indirect effects, but in every case where the total impact is significant at the 95% level of confidence, the indirect effects are themselves significant.

The 'spatial pv' is the p-value of a Wald test of the null hypothesis that spatial effects do not contribute to the explanatory power of the model. The 'model pv' is the standard Wald statistic for the overall explanatory power of the model.

Data

This study combines data from several sources: household survey data, aggregated to state level; daily confirmed case and death rates by state, aggregated as described below; and demographic data at the state level, as well as policy measures, from several sources.

US Household Pulse survey

We drew data from all available cohorts of the 2020-2021 Household Pulse Survey, conducted by the U.S. Census Bureau and other government agencies to track effects of COVID-19 on US residents. These data are publicly available through the website of the US Census Bureau.[1]

[1] https://www.census.gov/programs-surveys/household-pulse-survey/datasets.html

Pulse conducted on-line surveys with adults in American households across all 50 states and Washington, DC in weekly or biweekly cross-sectional samples drawn from April 23, 2020 through July 21, 2020 (Phase 1), August 19, 2020 through October 26, 2020 (Phase 2), October 28, 2020 through March 29, 2021 (Phase 3) and April 14, 2021 through July 5, 2021 (Phase 3.1). A small proportion of respondents repeated surveys for one or two additional weeks; we included each respondent's first survey only.

As the first twelve weeks (Phase 1) of the Household Pulse data were drawn from single weeks while all subsequent data are drawn from two-week periods, the data for the first twelve weeks was aggregated into six "waves". Waves 7 and 8 are considered missing data due to the one-month delay in starting Phase 2 of the Household Pulse survey. The first "week" from Phase 2 is labeled wave 9, and so on, with the latest Phase 3.1 data for "week" 33 as wave 29. In total, we have 29 waves of data, based on 2,510,501 household-wave observations.

The data contain sample weights which adjust for nonresponse and sampling stratification to produce estimates representative of the US adult population. These weights are used to collapse data on particular characteristics such as age group, race/ethnic category and gender into state-level average percentages for each wave. Thus, the Household Pulse data entering the models vary by state and wave.

USAFacts confirmed case and death rates

Using USAFacts[2] and state population statistics, we constructed a daily-frequency panel data set for the 48 contiguous U.S. states and the District of Columbia over the study period. This dataset of the confirmed case rates and COVID-19 death rates (per 100,000) was then aggregated by averaging into the two-week intervals corresponding to the 29 waves of the Household Pulse survey.

These tables show the confirmed case rate and death rate over the waves of the study. These are cumulative figures that illustrate the spread of the pandemic over the period of analysis.

[2] https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

USAFacts confirmed case and death rates

Using USAFacts[2] and state population statistics, we constructed a daily-frequency panel data set for the 48 contiguous U.S. states and the District of Columbia over the study period. This dataset of the confirmed case rates and COVID-19 death rates (per 100,000) was then aggregated by averaging into the two-week intervals corresponding to the 29 waves of the Household Pulse survey.

These tables show the confirmed case rate and death rate over the waves of the study. These are cumulative figures that illustrate the spread of the pandemic over the period of analysis.

[2] https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/

Table: Confirmed case rates per 100,000 by state

	Mean	SD	Min	Max
Wave 1 23 Apr 12 May	307.7	326.8	42.4	1593.9
Wave 2 14 26 May	443.3	406.5	44.5	1822.6
Wave 3 28 May 9 Jun	537.0	440.8	49.6	1921.8
Wave 4 11 23 Jun	627.4	447.0	61.0	1980.3
Wave 5 25 Jun 7 Jul	758.4	443.9	98.5	2027.8
Wave 6 9 21 Jul	954.6	473.9	197.2	2076.7
Wave 9 19 31 Aug	1549.2	669.0	252.2	3100.8
Wave 10 2 14 Sep	1724.2	712.2	265.8	3311.1
Wave 11 16 28 Sep	1915.6	757.3	275.9	3486.9
Wave 12 30 Sep 12 Oct	2146.4	815.2	290.8	3643.0
Wave 13 14 26 Oct	2457.8	928.2	316.1	4407.5
Wave 14 28 Oct 9 Nov	2919.8	1155.8	361.2	6314.9

	Mean	SD	Min	Max
Wave 15 11 23 Nov	3681.6	1542.8	500.4	8777.4
Wave 16 Nov 25 Dec 7	4541.1	1853.5	703.4	10496.3
Wave 17 9 21 Dec	5446.1	2022.5	950.7	11600.3
Wave 18 6 18 Jan	7105.4	2161.9	1480.9	12462.7
Wave 19 20 Jan 1 Feb	7806.5	2204.1	1807.7	12732.4
Wave 20 3 15 Feb	8273.1	2241.5	2098.1	12894.4
Wave 21 17 Feb 1 Mar	8568.6	2260.9	2350.9	13047.9
Wave 22 3 15 Mar	8843.8	2273.2	2643.3	13221.0
Wave 23 17 29 Mar	9042.1	2273.1	2884.7	13371.1
Wave 24 14 26 Apr	9624.1	2250.5	3546.8	13917.1
Wave 25 28 Apr 10 May	9838.2	2238.1	3715.7	14162.3
Wave 26 12 24 May	9990.7	2227.3	3835.7	14336.3
Wave 27 26 May 7 Jun	10084.1	2223.4	3881.9	14435.6
Wave 28 9 21 Jun	10144.2	2224.2	3900.4	14495.7
Wave 29 23 Jun 5 Jul	10197.8	2229.4	3910.9	14524.0

Table: Death rates per 100,000 by state

	Mean	SD	Min	Max
Wave 1 23 Apr 12 May	16.3	24.2	1.2	126.4
Wave 2 14 26 May	23.8	31.5	1.7	147.2
Wave 3 28 May 9 Jun	28.1	34.6	1.8	153.9
Wave 4 11 23 Jun	31.2	36.5	2.0	158.1
Wave 5 25 Jun 7 Jul	34.0	39.0	2.3	168.5
Wave 6 9 21 Jul	36.5	39.7	3.4	176.0
Wave 9 19 31 Aug	46.2	39.6	6.4	179.4
Wave 10 2 14 Sep	49.2	39.6	7.3	180.2
Wave 11 16 28 Sep	52.1	39.5	8.5	181.1
Wave 12 30 Sep 12 Oct	55.1	39.4	9.2	181.9
Wave 13 14 26 Oct	58.6	39.1	9.3	182.9
Wave 14 28 Oct 9 Nov	63.2	38.7	9.3	184.5

	Mean	SD	Min	Max
Wave 15 11 23 Nov	69.5	38.7	9.7	187.5
Wave 16 Nov 25 Dec 7	78.2	39.5	11.8	193.0
Wave 17 9 21 Dec	90.0	41.4	16.4	202.2
Wave 18 6 18 Jan	113.9	46.2	25.5	226.4
Wave 19 20 Jan 1 Feb	126.1	48.8	27.5	238.6
Wave 20 3 15 Feb	138.0	51.6	29.8	249.8
Wave 21 17 Feb 1 Mar	146.3	53.5	32.2	259.4
Wave 22 3 15 Mar	153.1	54.6	34.0	268.1
Wave 23 17 29 Mar	156.6	55.2	35.5	273.3
Wave 24 14 26 Apr	163.9	56.4	38.9	284.3
Wave 25 28 Apr 10 May	166.3	56.9	39.8	289.3
Wave 26 12 24 May	168.6	57.3	40.7	293.0
Wave 27 26 May 7 Jun	170.9	57.5	40.9	294.5
Wave 28 9 21 Jun	172.5	57.6	41.0	296.1
Wave 29 23 Jun 5 Jul	173.6	57.6	41.2	297.3

State-level and spatial data

Additional data included in our state-level models are also time-varying by wave. Measures of vaccine eligibility for various age groups were acquired from the 14 July 2021 release of the COVID-19 US state policies (CUSP) database available from https://statepolicies.com.

Spatial data, including the spatial unit identifier, geographic coordinates and geographic entity codes (GEOIDs) (i.e, the FIPS identifier) for each state were obtained from the US Census Bureau's TIGER geographic database.

State-level and spatial data

Additional data included in our state-level models are also time-varying by wave. Measures of vaccine eligibility for various age groups were acquired from the 14 July 2021 release of the COVID-19 US state policies (CUSP) database available from https://statepolicies.com.

Spatial data, including the spatial unit identifier, geographic coordinates and geographic entity codes (GEOIDs) (i.e, the FIPS identifier) for each state were obtained from the US Census Bureau's TIGER geographic database.

Household Pulse state/wave demographics include the distributions of survey respondents' age, race/ethnicity, education, and income. These measures were collapsed to the state level using the Household Pulse survey weights.

Table: Household Pulse: Age distribution

	Mean	SD	Min	Max	p25	p50	p75
% 18-29	0.172	0.034	0.068	0.293	0.151	0.172	0.191
% 30-39	0.189	0.027	0.124	0.358	0.172	0.185	0.201
% 40-49	0.166	0.016	0.108	0.265	0.156	0.165	0.175
% 50-59	0.168	0.018	0.113	0.237	0.157	0.168	0.179
% 60-69	0.178	0.023	0.101	0.255	0.163	0.179	0.193
% 70+	0.126	0.020	0.065	0.200	0.114	0.126	0.138

Table: Household Pulse: Race/ethnic distribution

	Mean	SD	Min	Max	p25	p50	p75
% White non-H	0.708	0.148	0.337	0.960	0.597	0.733	0.824
% Black non-H	0.104	0.099	0.001	0.491	0.028	0.070	0.141
% Asian non-H	0.033	0.027	0.000	0.157	0.015	0.024	0.042
% Other non-H	0.039	0.022	0.006	0.161	0.025	0.033	0.046
% Hispanic	0.116	0.098	0.013	0.526	0.051	0.084	0.133

Table: Household Pulse: Education distribution

	Mean	SD	Min	Max	p25	p50	p75
% HS or below	0.392	0.055	0.169	0.553	0.355	0.388	0.431
% Some college	0.312	0.039	0.158	0.502	0.287	0.310	0.340
% Bachelors degree	0.165	0.029	0.094	0.267	0.146	0.166	0.186
% Grad degree	0.131	0.042	0.065	0.417	0.105	0.122	0.143

Table: Household Pulse: Household income distribution

	Mean	SD	Min	Max	p25	p50	p75
% HH inc <35K	0.259	0.059	0.133	0.501	0.217	0.255	0.297
% HH inc 35-99K	0.452	0.046	0.241	0.598	0.425	0.455	0.485
% HH inc 100K+	0.288	0.072	0.130	0.562	0.238	0.274	0.332

Other demographics from Household Pulse include gender, family structure, the availability of health insurance, and whether medical care was delayed by the pandemic.

Table: Household Pulse: Other demographics

	Mean	SD	Min	Max	p25	p50	p75
% Female	0.514	0.013	0.453	0.588	0.506	0.515	0.521
% Single/adults only	0.613	0.037	0.486	0.734	0.589	0.614	0.637
% Family with kids	0.387	0.037	0.266	0.514	0.363	0.386	0.411
% No health ins	0.266	0.055	0.113	0.434	0.229	0.266	0.302
% Delayed medical care	0.293	0.091	0.085	0.494	0.232	0.304	0.361

Table: State-level regressors

	Mean	SD	Min	Max
% Vaccine elig 75+	0.377	0.485	0.000	1.000
% Vaccine elig 50+	0.224	0.417	0.000	1.000
% Vaccine elig gen public	0.216	0.412	0.000	1.000

Estimation results

In order to evaluate spatial heterogeneity, all models have been estimated for the continental US (49 states, including the District of Columbia) and separately for the four regions defined by the US Census Bureau.

Region	States
Northeast	9
South	17 (including DC)
Midwest	12
West	11

Modeling confirmed case rates

The first set of models for the COVID-19 confirmed case rate contain only the Household Pulse demographic variables. The omitted group contains 18-29 year old male respondents, either single or in an adults-only household, White non-Hispanic, with no more than a high school education and gross household income no more than $35,000. Those in the omitted group have health insurance and did not experience delays in medical care due to the pandemic.

Strong support for the spatial lags appears in all models. There is considerable variation of the average total impact measures across the regional estimates. The increasing impact of age on the confirmed case rate is evident. The lack of health insurance has a strong influence on the confirmed case rate, while delayed medical care reduces exposure, perhaps reflecting greater caution from patients and providers.

Table: Average Total Impact of spatial models of caserate

	US	Northeast	South	Midwest	West
% 30-39	21624.2***	10729.1	16078.2***	43787.2***	53469.9***
% 40-49	35313.6***	28278.9***	24605.4***	38963.8***	46026.0***
% 50-59	24430.7***	21922.3*	10733.5*	39381.5***	31635.2*
% 60-69	35706.3***	30969.6**	31292.1***	33918.6***	27748.9*
% 70+	34208.2***	28024.6**	28853.6***	31184.4***	83262.7***
% Family with kids	-1873.6	-9801.2	-112.7	-2242.4	-1063.3
% Black non-H	-1669.1	6479.7	-4645.9*	-7520.4	-18799.9
% Asian non-H	15558.1**	11456.4	805.1	-6619.3	16622.4
% Other non-H	-7043.1	-32105.0*	9712.5	-18551.6	-19748.0
% Hispanic	1898.4	7934.6	189.0	8215.5	5421.0
% Some college	14601.9***	-8995.4	8000.2	13576.2*	-9780.9
% Bachelors degree	5924.1	-24477.7	15291.9**	-10210.9	28277.8
% Grad degree	6617.0	-17035.3	14202.4**	-4857.6	19913.8
% Female	-510.9	12235.1	4826.3	-10022.5	-17933.8
% no health ins	18564.9***	21496.9***	14979.7***	11177.4**	30122.4***
% Delayed medical care	-24622.9***	-17144.9***	-24170.7***	-30874.7***	-20568.5***
% HH inc 35-99K	-8639.8***	-9774.5*	-3891.9	-3102.2	-12859.0*
% HH inc 100K+	730.2	4929.3	-2514.3	2435.1	-18391.2*
$\hat{\rho}$	0.508***	0.338***	0.222***	0.269***	0.662***
$\hat{\lambda}$	0.616***	0.535***	0.626***	0.671***	0.398***
LogLikelihood	-11405.952	-2122.889	-3982.900	-2829.820	-2510.832
PseudoR2	0.685	0.673	0.743	0.751	0.752
model pv	0.000	0.000	0.000	0.000	0.000
spatial pv	0.000	0.000	0.000	0.000	0.000

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

The second set of models augment the demographic variables with time-varying state level variables, capturing periods during which vaccines became available for those 75+ years old, those 50+ years old and for the general public. As the vaccine eligibility measures only appear in the last several waves, their positive values appear to be picking up trends in the recent confirmed case rate due to the Delta variant of the coronavirus.

Considerable differences in the impact measures appear across the US reghions.

The second set of models augment the demographic variables with time-varying state level variables, capturing periods during which vaccines became available for those 75+ years old, those 50+ years old and for the general public. As the vaccine eligibility measures only appear in the last several waves, their positive values appear to be picking up trends in the recent confirmed case rate due to the Delta variant of the coronavirus.

Considerable differences in the impact measures appear across the US reghions.

Table: Average Total Impact of spatial models of caserate

	US	Northeast	South	Midwest	West
% 30-39	9908.4***	603.9	6723.5**	22052.2***	30574.9***
% 40-49	17128.3***	12297.5*	11079.8***	20231.8***	22586.9**
% 50-59	16857.2***	17882.6**	9922.2**	30518.6***	22932.7**
% 60-69	22088.5***	21796.9***	17830.4***	27529.1***	25384.5***
% 70+	22371.3***	22413.6**	21597.3***	25950.0***	47705.1***
% Family with kids	339.4	-4847.7	2123.2	-344.9	-718.2
% Black non-H	-3203.0	10228.0	-4431.0*	-13070.1***	-14048.2
% Asian non-H	6510.2	12539.6	2141.4	-6577.7	5601.7
% Other non-H	-4830.9	-11016.4	3779.9	-13237.4*	-11601.3
% Hispanic	-630.0	-866.0	-1594.5	1086.4	3196.0
% Some college	8945.7***	2801.6	2157.4	5532.2	1381.9
% Bachelors degree	7083.9*	-1019.5	17620.9***	455.1	14789.8
% Grad degree	1230.4	-7976.4	7282.5	-3280.6	14453.3
% Female	2036.3	7822.1	5705.5	-2023.8	-11292.1
% no health ins	11358.6***	15121.0***	10836.5***	3021.5	19733.5***
% Delayed medical care	-3743.2***	4959.9*	-6138.7***	-6904.3***	-79.12
% HH inc 35-99K	-3550.1**	-1592.6	-1722.8	-1582.5	-6669.9
% HH inc 100K+	1278.9	9099.9**	-3155.5	1448.2	-8524.6
% Vaccine elig 75+	4063.5***	4627.8***	3999.9***	5272.7***	4377.0***
% Vaccine elig 50+	1156.4***	335.8	710.9	1113.1	1944.2*
% Vaccine elig gen public	420.6	1418.6	636.1	-367.0	-452.5
$\hat{\rho}$	0.266***	0.123*	0.106**	-0.00675	0.486***
$\hat{\lambda}$	0.750***	0.553***	0.640***	0.812***	0.615***
LogLikelihood	-11225.183	-2071.505	-3876.032	-2778.686	-2478.158
PseudoR2	0.748	0.758	0.841	0.774	0.782
model pv	0.000	0.000	0.000	0.000	0.000
spatial pv	0.000	0.000	0.000	0.000	0.000

Modeling death rates

We now turn to models of COVID-19 state/wave death rates. These models differ from the confirmed case models by including the prior wave's confirmed case rate as an explanatory factor, reflecting the likelihood that some cases of COVID-19 will result in death of the infected patient, keeping in mind that improvements in medical care have reduced the likelihood that an infected patient will be hospitalized or fail to survive.

In these models of demographic factors, there is a significant effect of the lagged case rate across the board, with similar magnitudes. The Northeast has the largest total impact value, perhaps reflecting the early peak of the pandemic in those states. Western states have the lowest value. The model of Northeast states fits more closely than those of other regions, but there is weak evidence of spatial effects in that model.

Table: Average Total Impact of spatial models of death rates

	US	Northeast	South	Midwest	West
lcaserate	0.0141***	0.0151***	0.0144***	0.0141***	0.0125***
% 30-39	-1.763	-124.0***	49.52	-36.85	34.86
% 40-49	48.93	37.09	103.4**	-122.1**	9.038
% 50-59	92.83**	146.6**	125.8**	8.023	6.992
% 60-69	94.31**	99.04	98.74*	-42.31	165.0***
% 70+	142.8***	149.7**	110.9*	-29.35	350.1***
% Family with kids	-42.66**	33.88	-33.21	-48.93*	-62.01*
% Black non-H	-19.14	49.01	103.7**	-47.15	61.82
% Asian non-H	87.10	-7.803	89.96	-131.1	39.74
% Other non-H	68.24	199.0*	183.8**	-30.95	17.06
% Hispanic	9.857	-161.6**	76.79*	16.61	56.78*
% Some college	-66.01*	-31.28	-128.8*	52.24	-173.6**
% Bachelors degree	-96.10*	145.5*	-18.89	-336.2***	-173.5**
% Grad degree	-73.39	91.31	-57.02	-316.5***	14.86
% Female	127.7**	311.0***	20.05	166.1**	40.52
% no health ins	11.33	85.31***	7.092	13.94	-0.321
% Delayed medical care	0.115	98.80***	-60.31***	-32.83***	1.062
% HH inc 35-99K	46.29***	17.44	41.18	49.81*	28.18
% HH inc 100K+	57.71***	10.25	52.39	24.62	61.01
$\hat{\rho}$	0.0234	0.0530	0.146***	-0.113***	-0.349***
$\hat{\lambda}$	0.211***	-0.265**	0.0528	-0.327***	0.0927
LogLikelihood	-5080.665	-854.078	-1678.249	-1173.720	-1162.347
PseudoR2	0.701	0.677	0.885	0.824	0.894
model pv	0.000	0.000	0.000	0.000	0.000
spatial pv	0.000	0.029	0.000	0.000	0.000

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

The second set of models for the state/wave death rate augment the demographic variables with three time-varying state level variables, capturing periods during which vaccines became available for those 75+ years old, those 50+ years old and the general public. Positive, significant coefficients on the variable denoting eligibility to the oldest cohort are puzzling, as those states more rapid deployment of the vaccine might be expected to exhibit more favorable outcomes.

Table: Average Total Impact of spatial models of death rates

	US	Northeast	South	Midwest	West
lcaserate	0.0137***	0.0164***	0.0146***	0.0137***	0.0117***
% 30-39	0.900	-90.38**	47.41	-10.67	42.47
% 40-49	51.68	20.15	109.2**	-109.9*	2.198
% 50-59	106.4***	88.08	145.6***	48.47	30.26
% 60-69	108.0***	37.70	110.5*	-22.71	207.7***
% 70+	158.8***	93.16	137.6**	-6.041	388.9***
% Family with kids	-41.09**	34.10	-29.44	-47.09*	-55.62
% Black non-H	-30.13	52.33	97.93**	-42.58	19.74
% Asian non-H	78.70	-22.74	85.21	-129.0	39.77
% Other non-H	62.86	201.2**	165.8**	-28.19	19.72
% Hispanic	6.631	-123.1	77.33*	7.391	53.11
% Some college	-64.71*	-54.84	-120.6*	40.75	-149.3**
% Bachelors degree	-86.05	91.90	-14.91	-297.4***	-148.1*
% Grad degree	-73.24	62.45	-39.89	-268.8***	9.909
% Female	133.3***	307.2***	17.99	171.8**	32.07
% no health ins	15.57	71.92**	7.864	17.35	17.12
% Delayed medical care	13.61	40.25**	-40.62*	-20.64	46.22**
% HH inc 35-99K	48.37***	4.549	42.46	48.68*	38.21
% HH inc 100K+	59.22***	-17.74	48.00	21.36	67.17*
% Vaccine elig 75+	3.521*	-6.776**	-2.599	5.964***	8.689**
% Vaccine elig 50+	1.712	-1.897	11.68	8.714	-5.487
% Vaccine elig gen public	0.248	-10.28	-7.054	-9.883	11.96
$\hat{\rho}$	0.0205	0.0606*	0.152***	-0.122***	-0.356***
$\hat{\lambda}$	0.200***	-0.377***	0.0398	-0.388***	0.0889
LogLikelihood	-5078.748	-843.971	-1675.901	-1168.114	-1156.652
PseudoR2	0.696	0.724	0.884	0.834	0.892
model pv	0.000	0.000	0.000	0.000	0.000
spatial pv	0.000	0.001	0.000	0.000	0.000

Concluding remarks

This preliminary investigation into COVID-19 modeling with spatial panel autoregressions illustrates the usefulness of incorporating the time dimension into modeling the pandemic's spread. The evolution of COVID-19 in the US over the last 14 months has proceeded at different rates across the country, as influenced by policy decisions as well as demographic factors.

These models are a first step toward capturing those spatial impact effects and their geographic heterogeneity. In further work, we are testing a number of other factors and model specifications, such as fully heterogenous spatial autoregressive panel models, to strengthen our findings.