A Two-Stage Time Series Clustering Framework for Explaining the Varying Patterns of COVID-19 Deaths across the U.S.

Fadel M. Megahed, L. Allison Jones-Farmer, Yinjiao Ma, Steven Rigdon

Submitted to: JMIR Public Health and Surveillance
on: July 28, 2021

Table of Contents

Original Manuscript **4**
Supplementary Files **21**
Figures **22**
Figure 1 ..... **23**
Figure 2 ..... **24**
Figure 3 ..... **25**
Figure 4 ..... **26**
Figure 5 ..... **27**
Figure 6 ..... **28**

A Two-Stage Time Series Clustering Framework for Explaining the Varying Patterns of COVID-19 Deaths across the U.S.

Fadel M. Megahed[1] PhD; L. Allison Jones-Farmer[1] PhD; Yinjiao Ma[2] MPH; Steven Rigdon[2] PhD

[1]Farmer School of Business Miami University Oxford US
[2]Department of Epidemiology and Biostatistics College for Public Health and Social Justice Saint Louis University Saint Louis US

Corresponding Author:
Steven Rigdon PhD
Department of Epidemiology and Biostatistics
College for Public Health and Social Justice
Saint Louis University
3545 Lafayette Ave
Saint Louis
US

Abstract

Background: Socially vulnerable communities are at an increased risk for adverse health outcomes during a pandemic. While this association has been established for H1N1, MERS and COVID-19 outbreaks, understanding the factors influencing the outbreak pattern for different communities remains limited.

Objective: Our three objectives are to determine how many distinct clusters of time series there are for COVID deaths in the 3,108 counties in the contiguous US, how the clusters are geographically distributed, and what factors influence the probability of cluster membership.

Methods: We propose a two-stage data analytic framework that can account for different levels of temporal aggregation for the pandemic outcomes and community-level predictors. Specifically, we use time-series clustering to identify clusters with similar outcome patterns for the 3,108 contiguous U.S. counties. Multinomial logistic regression is used to explain the relationship between community-level predictors and cluster assignment. We analyzed county-level confirmed COVID-19 deaths from Sunday March 1, 2020 to Saturday February 27, 2021.

Results: Four distinct patterns of deaths were observed across the contiguous U.S. The multinomial regression model correctly classified 61.25\% of the counties' outbreak patterns/clusters.

Conclusions: Our results provide evidence that county-level patterns of COVID-19 deaths are different, and can be explained in part by social and political predictors.

(JMIR Preprints 28/07/2021:32164)
DOI: https://doi.org/10.2196/preprints.32164

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
✓ **Only make the preprint title and abstract visible.**
No, I do not wish to publish my submitted manuscript as a preprint.
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

A Two Stage Time Series Clustering Framework for Explaining the Varying Patterns of COVID-19 Deaths across the U.S.

Abstract

Background: Socially vulnerable communities are at an increased risk for adverse health outcomes during a pandemic. While this association has been established for H1N1, MERS and COVID-19 outbreaks, understanding the factors influencing the outbreak pattern for different communities remains limited.

Objectives: Our three objectives are to determine how many distinct clusters of time series there are for COVID deaths in the 3,108 counties in the contiguous US, how the clusters are geographically distributed, and what factors influence the probability of cluster membership.

Methods: We propose a two-stage data analytic framework that can account for different levels of temporal aggregation for the pandemic outcomes and community-level predictors. Specifically, we use time-series clustering to identify clusters with similar outcome patterns for the 3,108 contiguous U.S. counties. Multinomial logistic regression is used to explain the relationship between community-level predictors and cluster assignment. We analyzed county-level confirmed COVID-19 deaths from Sunday March 1, 2020 to Saturday February 27, 2021.

Results: Four distinct patterns of deaths were observed across the contiguous US. The multinomial regression model correctly classified 61.25% of the counties' outbreak patterns/ clusters.

Conclusions: Our results provide evidence that county-level patterns of COVID-19 deaths are different, and can be explained in part by social and political predictors.

Keywords: explanatory modeling; multinomial regression; SARS-CoV-2; socio-economic analyses; time-series analysis

Introduction

A geographically, politically, and socioeconomically diverse nation, the US consists of fifty states, forty-eight of which are contiguous. When considering the COVID-19 pandemic in different regions throughout the US, different patterns of outcomes emerge. Based on data obtained from the open-source COVID-19 data hub [1], Figure 1 shows the national seven-day moving average of deaths as well as the various patterns that arise among eight example counties from Sunday, March 1, 2020 to February 27, 2021. For example, New York, NY experienced a large first wave of deaths, followed by a relatively low death count through the remainder of the study. Nearby Ocean County, NJ, a populous county near the New Jersey shore had a large first wave of deaths followed by a second wave beginning in late 2020. On the other hand, Butler County, OH, a populous Midwestern county showed very low death counts until late in the study period. None of these patterns mimics the overall pattern for the aggregate death counts in the U.S.

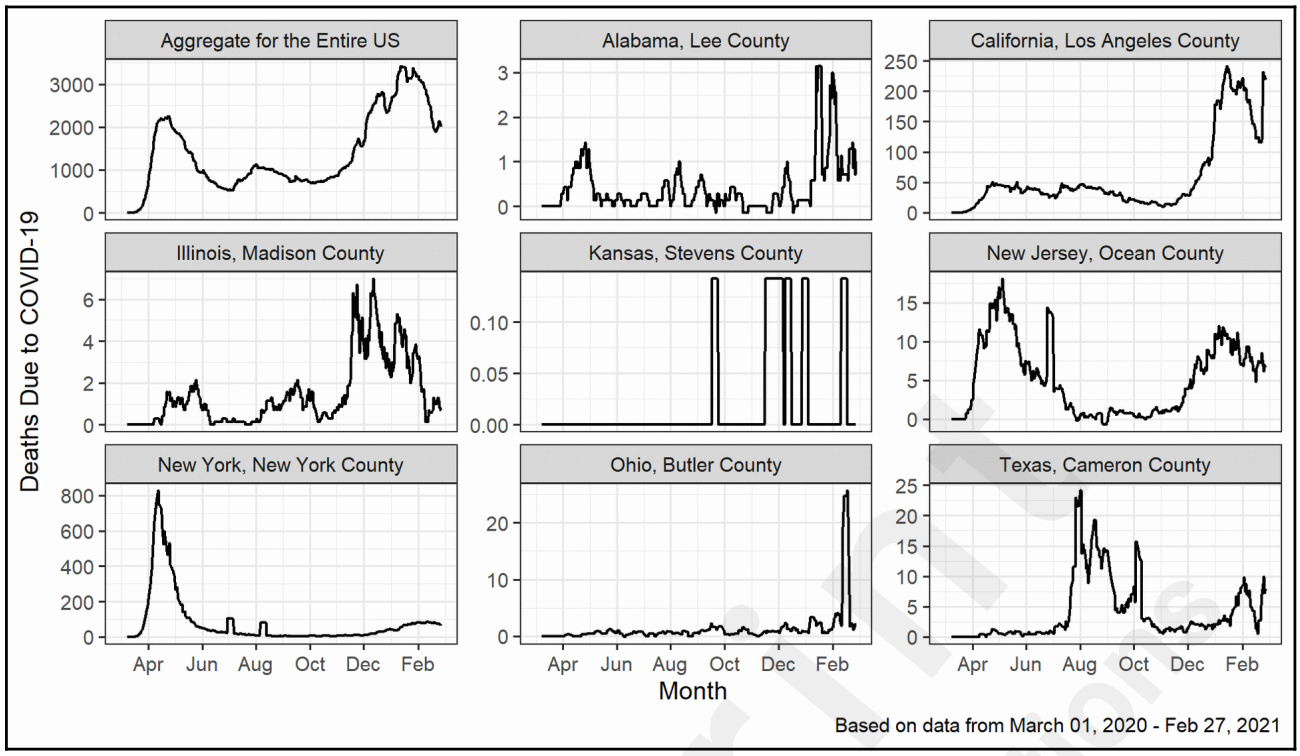


Figure 1. The time series profiles of the 7-day moving average of new COVID-19 deaths for the entire US and eight sample counties.

Early in the COVID-19 pandemic, the county-level population mortality and case fatality rates were significantly different among the US regions [2]. Explanations for regional differences in health outcomes related to COVID-19 may be the structure of the government and policy making within the US as it relates to the social vulnerability of the population. In the US, each state consists of county governments that set health and economic policies for local communities. The counties within the states vary in terms of population size, demographics, access to health care, housing, and transportation. Some have noted that the regional differences in COVID-19 policies, compliance and subsequent outcomes could be due to political differences across the regions. Goldwitzer et al. [3] showed Republican-leaning counties displayed less physical distancing compared to Democratic-leaning counties and a subsequent increase in COVID-19 cases and deaths. Another study showed Democratic governors were 50% more likely to implement stay-at-home orders [4], which have been associated with increased physical distancing and reduction of COVID-19 cases and deaths [5].

Here, we investigate the regional patterns in deaths attributed to COVID-19. The phenomenon of differing national and regional patterns within the US was illustrated for confirmed COVID-19 cases in [6]. In addition, a report by the *Financial Times* [7] argued “Across the world, public health data are gathered at a very local level before aggregation into regional and national figures.... While useful as a summary, local distinctions get lost, painting a misleading image of whole countries being affected uniformly.” In this study, we investigate the various patterns of COVID-19 deaths across the 3,108 counties in the contiguous United States. We also seek to determine what factors relate to the pattern of deaths. Specifically, we pose three questions:

1. How many distinct clusters of counties in the US exhibit similar time series patterns in the deaths due to COVID-19?
2. How are these clusters geographically distributed across the US?
3. Are certain geographic, political, government, and social vulnerability variables associated with the patterns of COVID-19 related deaths?

To address the first question, we performed a cluster analysis on the time series of the 3,108 US counties. We provide maps to show the geographic distribution of the clusters. To address the third question, we applied a multinomial logistic regression analysis using geographic, political, and social vulnerability data to explain the patterns of deaths due to COVID-19 over time.

Methods

This study was conducted in three stages: (1) data gathering and preprocessing; (2) time series clustering; and (3) modeling and cluster validation.

Data

The open-source COVID-19 data hub [1] was used to extract county-level time series data related to confirmed COVID-19 deaths from Sunday, March 1, 2020 to February 27, 2021. Data were extracted from 3,108 counties in the 48 contiguous US and were completely anonymous. This data set was used to compute the daily confirmed deaths related to COVID-19 by county and was the sole data used to inform the time series cluster analysis.

To develop the explanatory model describing the clusters, additional variables were gathered including the following:

- **Region:** The CDC produces a ten-region Framework for Chronic Disease Prevention and Health Promotion [8]. Figure 2 shows the ten regions used in our explanatory model. The CDC's National Center for Chronic Disease Prevention and Health Promotion (NCCDPHP) developed these regions to promote consistency in technical assistance and communications for chronic disease prevention [8].
- **Governor's Party Affiliation:** The political party affiliation of each US state governor (within the 48 contiguous US States) at the start of the Pandemic (March 2020) was determined. Since the District of Columbia does not have a governor, the political party of the mayor (Democrat) was used. Party affiliation of the governor was used as this affects the political actions and policies taken, often in the form of executive orders from the governor, during the pandemic [4].
- **Government Response:** The overall government response index (at the US state level) from the Blavatnik School of Government [9] was downloaded on March 16, 2021. The index considers containment and closure indicators such as school and workplace closings; Economic Response such as income support and debt relief; and Health Systems such as testing policies, contact tracing and investment in vaccines. Higher values of the Government Response Index indicate stronger government response related to the pandemic. This index changes over the time of the study period. To capture the index over the majority of the study period, we summarized the index using the median value over the study period. Details on the methodology used to compute the index can be found at Oxford University COVID-19 Tracker Github [10].
- The Center for Disease Control and Prevention's (CDC) **Social Vulnerability Index (SVI)** computed by the CDC's Agency for Toxic and Disease Registry's Geospatial Research, Analysis and Service program [11]. The SVI provide the relative vulnerability of each US county based on US Census data and are ranked on 15 social factors including unemployment, minority status, and disability. Note that the SVI data from the CDC returned results for 3,107 counties, with no data on Rio Arriba County, New Mexico, and hence this county will be excluded from our explanatory analysis. The SVI are grouped into four themes including the following:
    - RPL Theme1: Socioeconomic
    - RPL Theme2: Household Composition & Disability

- RPL Theme3: Minority Status & Language
- RPL Theme4: Housing & Transportation

Our study included each of the four SVI themes. To construct the SVI for each theme, the percentile rank for each variable across the counties is computed. These are summed across the themes, then ranked within each domain. The SVIs range from 0 to 1 with higher values of SVI for a particular theme indicating a higher-level of social vulnerability. For more details on the SVI, see Flanagan et al. [12].

- **Population Density:** The population density in each county was computed based on the land area in square miles and the 2014-2018 ACS (American Community Survey) population estimates in each county. Both the land area and population estimate variables were obtained from the CDC's Social Vulnerability Index 2018 dataset [11]. Due to right-skewness in this variable, the natural logarithm of population density is used in the analysis.

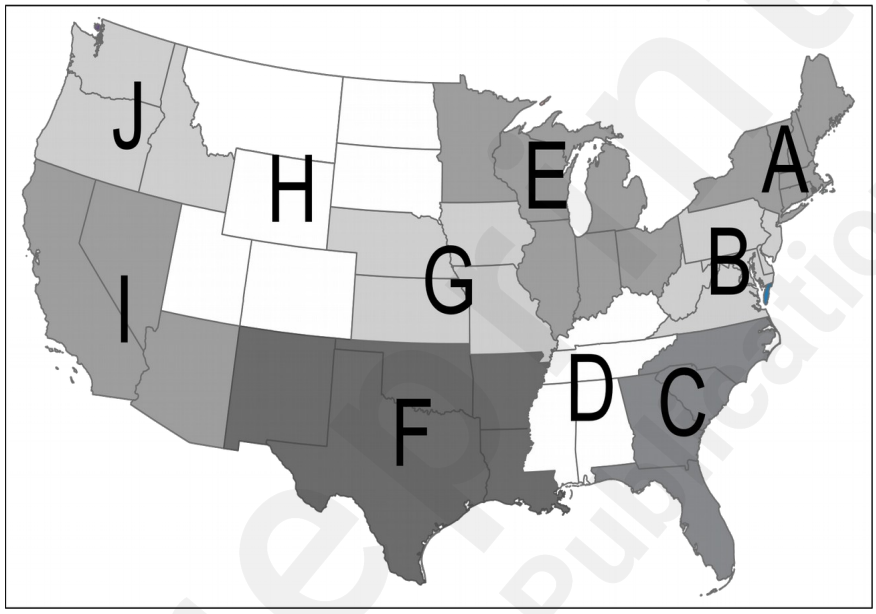


Figure 2. The ten CDC regions.

Time Series Clustering

Time series cluster analysis was based solely on the daily confirmed deaths related to COVID-19 by county. The goal is to separate counties into groups (clusters) that show similar time series patterns. There are three important decisions that affect the cluster solution including (1) the scaling of the data; (2) the measure of distance between the clusters; and (3) the clustering algorithm. Liao [13] gives an overview of time series clustering methods.

For this study the daily confirmed deaths related to COVID-19 by county were smoothed using a seven-day moving average to account for weekly patterns due to reporting. Moreover, the seven-day moving averages were rescaled so that all values fall between 0 and 1 to focus on the pattern of the progression of the deaths rather than the magnitude of the death counts. The magnitude of the death counts in each county depends on many factors such as county size, population density, region, etc. The scaled seven-day moving averages for county at time is

$$MA7_i^{(\text{scaled})} = \max\left(0, \frac{MA7_i}{\max(\boldsymbol{MA7})}\right)$$ where, is the seven-day moving average of deaths related to COVID-19 for county at a time . The maximum in the denominator is taken over all time, . The outer max function in (1) is used to account for reporting adjustments that occur with negative death

counts on some days.

For illustation, suppose that county recorded deaths only on days 7, 8, and 9, when respectively, 7, 21, and 14 deaths occurred. On all other days, no deaths were recorded. For clarity, this sequence of death counts, the calculations of the seven-day moving averages, , and the scaled moving averages, , for the first 17 days are shown in Table 1.

Table 1. Example calculation of the scaled seven-day moving averages .

Time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Deaths	0	0	0	0	0	0	7	21	14	0	0	0	0	0	0	0	0
	NA	NA	NA	NA	NA	NA	1	4	6	6	6	6	6	5	2	0	0
	NA	NA	NA	NA	NA	NA			1	1	1	1	1			0	0

This method of scaling the seven-day moving averages assures that we evaluate the shape of the death profile for each county across time.

Many metrics can be used to measure the distance between time series including Euclidean distance, dynamic time warping [14], Pearson's correlation coefficient, etc. An elastic measure such as dynamic time warping is commonly used with time series clustering [13] because it aligns or *warps* the time series so that the distance between them is minimized. Elastic measures such as this do not preserve the timing of the outbreak and deaths in a meaningful way. For this reason, we used Euclidean distance to measure the distance between the time series clusters. In our case, the Euclidean distance between two death profiles of length T is

$$d_E\left(MA7_i^{scaled}, MA7_j^{scaled}\right) = \sqrt{\sum_{t=1}^{T}\left(MA7_{i,t}^{scaled} - MA7_{j,t}^{scaled}\right)^2}$$

$$d_E(r,s) = \sqrt{\sum_{t=1}^{T}(r_t - s_t)^2}$$

There are numerous clustering algorithms that have been suggested for time series clustering [13,15]. We used k-means clustering for this analysis. A heuristic-based method of clustering, k-means partitions n objects into $k \leq n$ mutually exclusive clusters and each cluster is represented by the most centrally located object in a cluster. One limitation of the k-means clustering approach is that the number of clusters must be determined a priori in order to obtain a solution. It is common practice in exploratory research to evaluate cluster solutions for several sizes of k, and select the *best* based on measures of cluster validity or homogeneity [16]. The **R** package, *NbClust* [17] can be used to compute up to 30 cluster validity indices for cluster solutions of several sizes, k. This approach provides a systematic, data-driven method for selecting the optimal number of clusters in a data set without capitalizing on a single validity measure. For this analysis, *k*-means clustering was used to find the cluster solutions and the package *NBClust* was used to determine the optimal number of clusters to retain.

Explanatory Modeling

The time series clustering method described above results in mutually exclusive clusters of time series profiles containing counties with similar patterns in the daily deaths related to COVID-19. To further validate the cluster solution and to explain the differences in the progression of daily deaths across the counties, a multinomial regression analysis [18] was fit using the explanatory variables described in the *Data* subsection. The function *multinom* from the **R** package *nnet* [19] was used for this analysis.

Model performance is evaluated in terms of the ability to meaningfully interpret the model coefficients and by evaluating the in-sample classification performance. Specifically, the model predicted cluster is compared to the cluster as determined by the time series cluster solution for each county. The in-sample classification performance is measured by sensitivity, specificity, and balanced accuracy:

$$\text{sensitivity} = \frac{TP}{TP + FN},$$

where TP and FN are the number of true positive and false negative predictions,

$$\text{specificity} = \frac{TN}{TN + FP},$$

where TN and FP are the number of true negatives and false positive predictions, and

$$\text{balanced accuracy} = \frac{\text{sensitivity} + \text{specificity}}{2}.$$

Results

Number of Distinct Clusters

To address our first research question regarding the number of distinct clusters, we used time series cluster analysis of the scaled seven-day moving average of daily deaths due to COVID-19. Figure 3 shows the scaled time series of the daily deaths due to COVID-19 for nine randomly selected counties in the contiguous US during the study period. We evaluated $2 \leq k \leq 51$ time series cluster solutions using twenty-three cluster validity indices [17]. Seven of the 23 validity indices preferred a four-cluster solution. The second most preferred cluster solution is a two-cluster solution which was preferred by six of the 23 indices. Using a majority rule of the validity indices, we retained a four-cluster solution.

Figure 4 shows the geographic distribution of the four-cluster solution across the US. Cluster C1 is primarily concentrated in the Upper Midwest and mountain states, as well as in Ohio, central Kentucky, Virginia, and Maine. Cluster C2 is located along the coast in the Northeast, and in some of the larger US cities, such as Chicago, Detroit, Seattle, and New Orleans. Cluster C3 is scattered throughout much of the US, but particularly in Missouri, Illinois, and the states surrounding the Great Lakes. Cluster C4 occurs across the U.S., but shows concentrations in California, east Texas, the Southwest, and the Southeast.

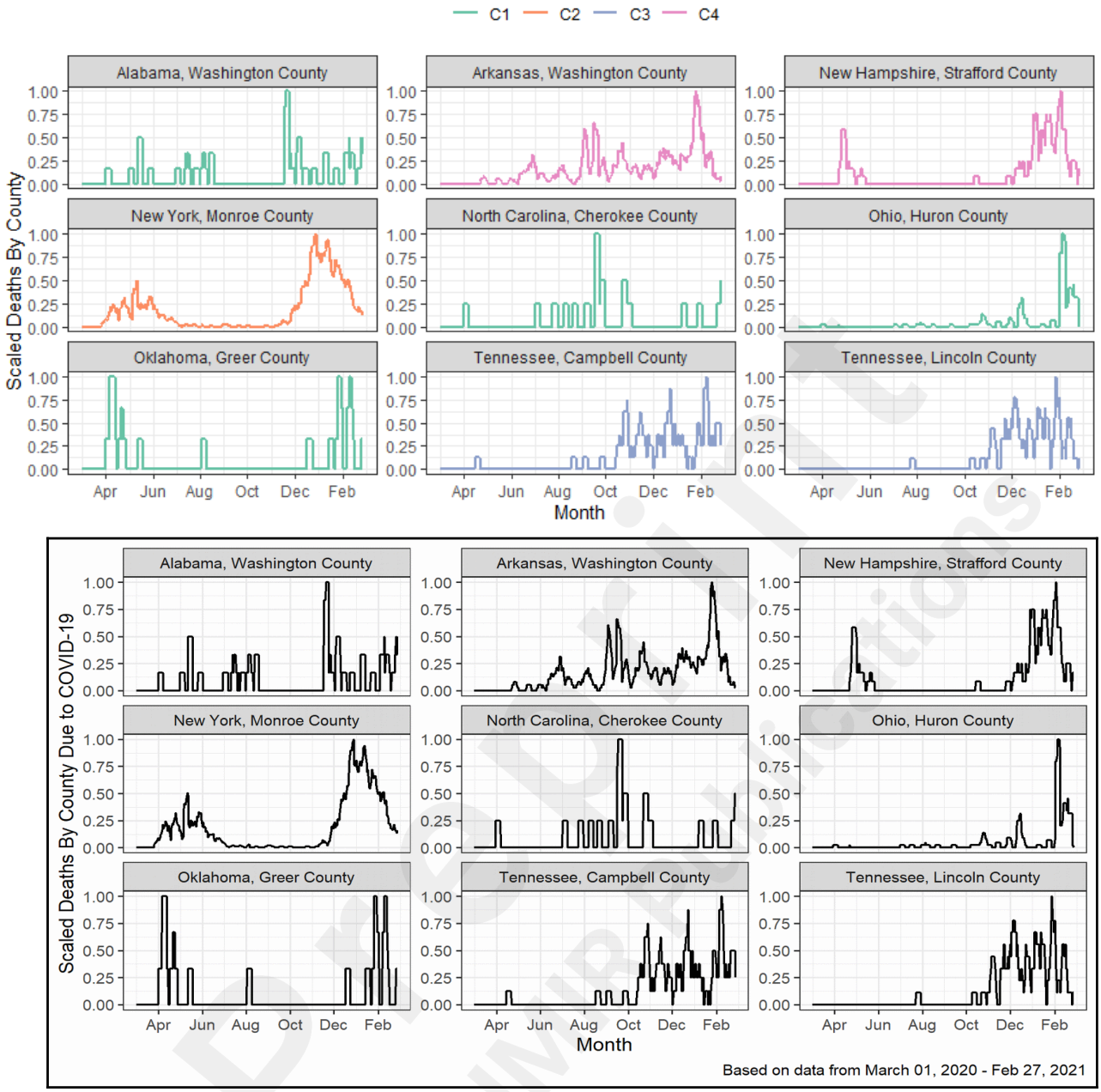


Figure 3. The time series profiles of the scaled 7-day moving average of new COVID-19 deaths for nine sample counties.

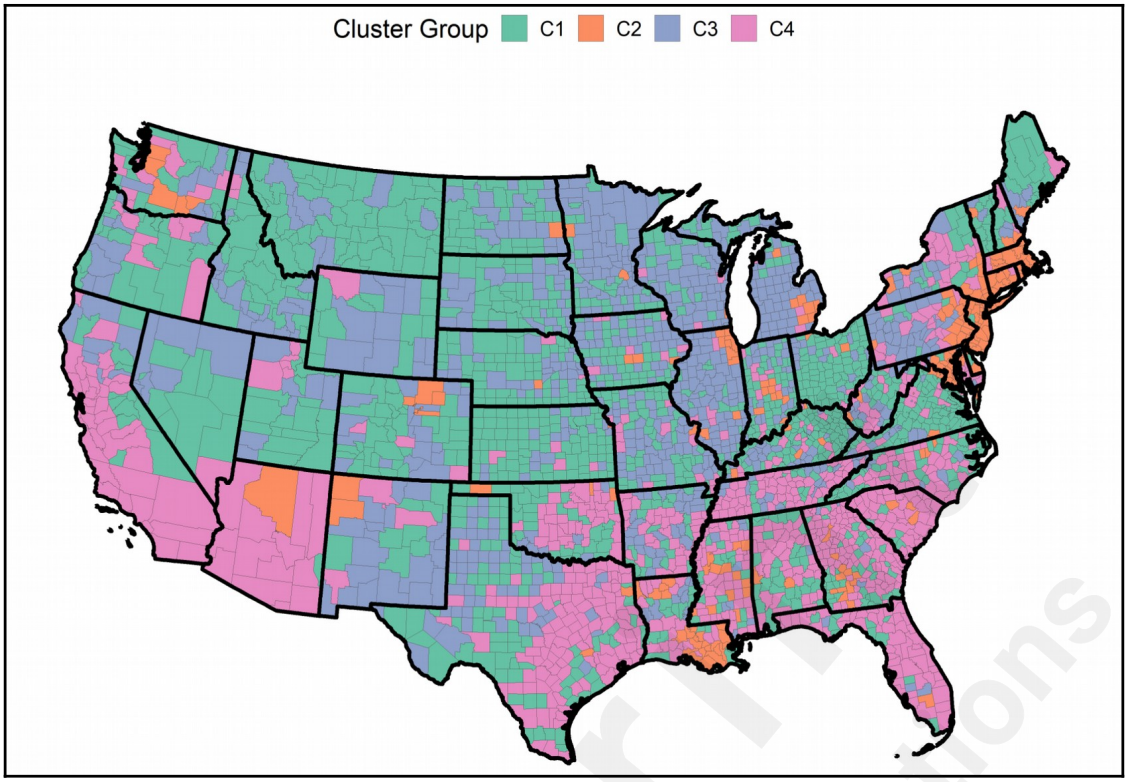


Figure 4. Map of the four scaled time-series profile clusters of COVID-19 deaths by county in the contiguous United States. For an interactive color version of this map, please see Section 3.3.3 of [20].

Figure 5 shows the 25th, 50th, and 75th percentiles of the time series profiles for the counties within each cluster and gives insight into the shape of the cluster patterns. From Figure 5, it is clear that counties in cluster C1 experienced a low number of deaths due to COVID-19 throughout the study period. Counties clustering in C2 experienced early death counts beginning in April 2020, but tapered off in early summer. These counties maintained low death counts throughout the late summer and early fall, until rising again in November 2020. In C3, counties experienced few COVID-19 deaths until October 2020, when they saw a rapid rise in deaths. The death counts in C3 began dropping in December 2020 which continued through March 2021. The fourth cluster, C4, showed a small increase in deaths in late summer, followed by a steady rise throughout the fall, and a higher peak in early 2021.

Explaining the Clusters

To address the second research question regarding factors that relate to the patterns of COVID-19 related deaths, we used an explanatory multinomial regression analysis to validate our cluster solution. Table 2 provides a summary of the explanatory study variables for each cluster.

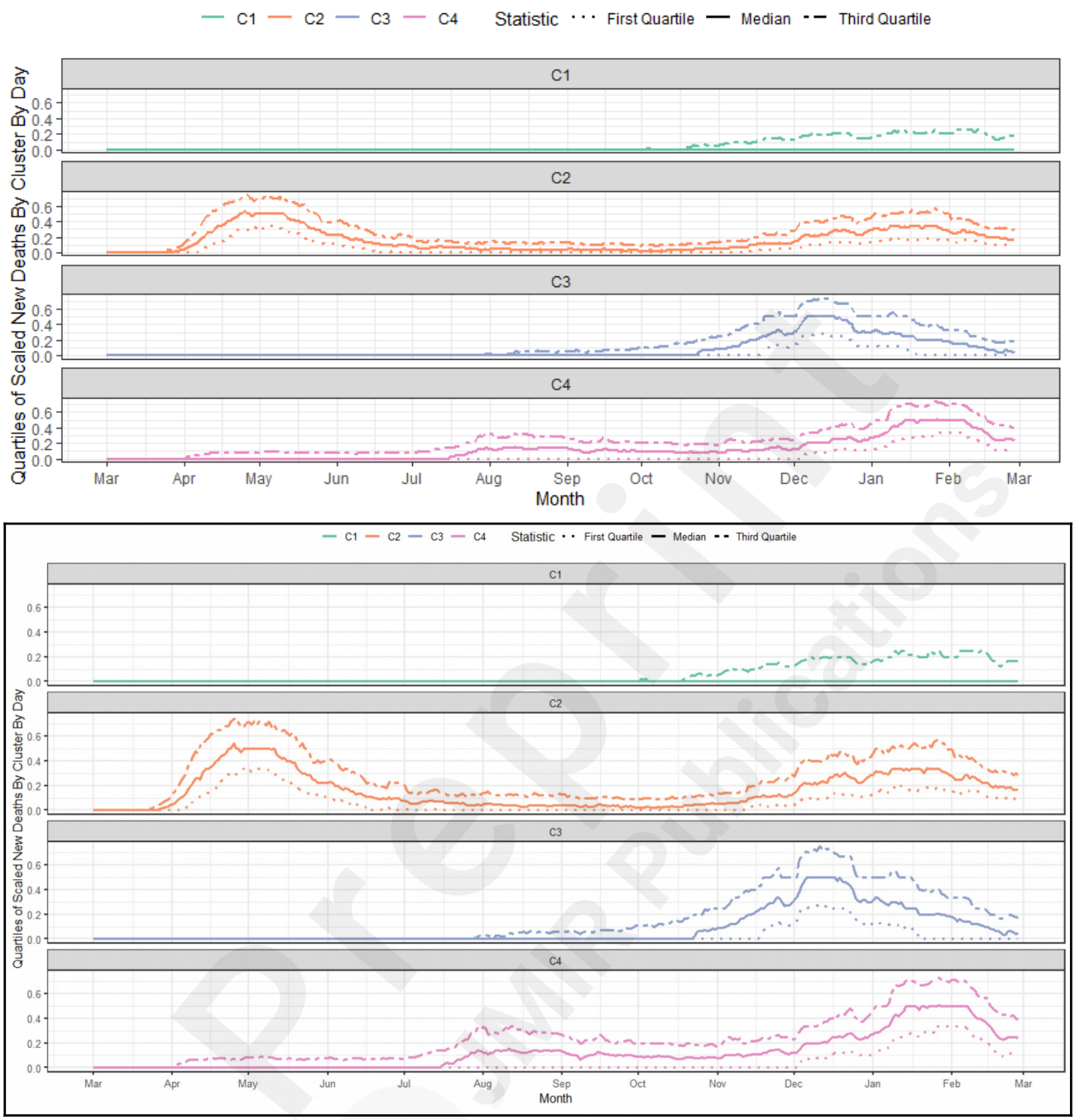


Figure 5. A summary plot, where the median scaled time-series profile for each cluster is depicted using the 'solid' **bold** line. The first and third quartiles are shown through 'dotted' and 'two-dash' lines, respectively.

Table 2: A summary of how the predictor variables are distributed per cluster. For each numeric variable, we report the mean ± standard deviation (SD). For categorical variables, we report the distribution of each subcategory across the four clusters. The row summation of percentages for a subcategory may deviate slightly from 100% due to rounding errors.

Continuous Variables	C1 (*n* = 1261) Mean ± SD	C2 (*n* = 226) Mean ± SD	C3 (*n* = 827) Mean ± SD	C4 (*n* = 794) Mean ± SD
Theme 1: Socioeconomic	0.48 ± 0.30	0.44 ± 0.31	0.45 ± 0.27	0.61 ± 0.26
Theme 2: Household Comp. & Disability	0.50 ± 0.28	0.37 ± 0.31	0.49 ± 0.28	0.56 ± 0.29
Theme 3: Minority Status & Language	0.41 ± 0.28	0.71 ± 0.22	0.43 ± 0.27	0.65 ± 0.24
Theme 4: Housing & Transportation	0.42 ± 0.29	0.60 ± 0.28	0.49 ± 0.26	0.60 ± 0.27
log(*Population Density*)	3.01 ± 1.71	5.86 ± 1.81	3.73 ± 1.31	4.60 ± 1.29
Government Response Index Median	47.09 ± 8.45	52.87 ± 9.13	47.24 ± 8.25	48.13 ± 7.65
Categorical Variables	**C1 (*n* = 1261) *n* (%)**	**C2 (*n* = 226) *n* (%)**	**C3 (*n* = 827) *n* (%)**	**C4 (*n* = 794) *n* (%)**
Governor's Party				
Democratic	579 (45.9%)	142 (62.8%)	428 (51.8%)	202 (25.4%)
Republican	682 (54.1%)	84 (37.2%)	399 (48.2%)	591 (74.4%)
Region				
A	41 (3.3%)	43 (19.0%)	21 (2.5%)	24 (3.0%)
B	131 (10.4%)	63 (27.9%)	62 (7.5%)	48 (6.0%)
C	101 (8.0%)	19 (8.4%)	13 (1.6%)	239 (30.1%)
D	140 (11.1%)	20 (8.8%)	51 (6.2%)	153 (19.3%)
E	188 (14.9%)	30 (13.3%)	283 (34.2%)	23 (2.9%)
F	154 (12.2%)	31 (13.7%)	116 (14.0%)	201 (25.3%)
G	236 (18.7%)	7 (3.1%)	144 (17.4%)	25 (3.1%)
H	187 (14.8%)	7 (3.1%)	88 (10.6%)	9 (1.1%)
I	22 (1.7%)	1 (0.4%)	14 (1.7%)	53 (6.7%)
J	61 (4.8%)	5 (2.2%)	35 (4.2%)	18 (2.3%)

Rio Arriba County, New Mexico, assigned to C4 based on the time-series clustering was not modeled using the multinomial logistic regression since we could not obtain values for its predictor variables. Hence, the reported mean ± standard deviations and *n* (%) for C4 exclude this county.

Table 3 gives the coefficients from the multinomial logistic regression analysis. The baseline category for the analysis was C1, the cluster of counties with very few deaths related to COVID-19. The coefficients show the linear change in the natural log of the odds ratio of a county classifying in a corresponding cluster (e.g. C2, C3, or C4) vs. the baseline cluster (C1). From Table 3, it is clear that several geographic, political, government and social vulnerability variables are associated with the patterns in COVID-19 related deaths.

Table 3. Results of multinomial logistic regression for classes C2, C3 and C4. We have used C1 as the reference cluster since it contained the largest number of counties.

	Dependent Variable (Cluster)					
	C2		C3		C4	
	$\hat{\beta}$ (std. error)	OR (95% CI)	$\hat{\beta}$ (std. error)	OR (95% CI)	$\hat{\beta}$ (std. error)	OR (95% CI)
Th1 Socioeconomic	0.419 (0.592)	1.52 (0.48, 4.85)	−¿ 0.356 (0.286)	0.70 (0.40, 1.23)	−¿ 0.018 (0.276)	0.98 (0.47, 2.05)
Th2 Household/Disability	−¿ 0.245 (0.422)	0.78 (0.34, 1.83)	0.392 (0.223)	1.48 (0.96, 2.29)	0.638 (0.267)	1.89 (1.12, 3.19)
Th3 Minority/Language	3.661 (0.469)	38.90 (15.51, 97.54)	0.004 (0.222)	1.00 (0.65, 1.55)	1.162 (0.268)	3.20 (1.89, 5.40)
Th4 Housing/Transport	0.557 (0.428)	1.75 (0.75, 4.04)	1.086 (0.227)	2.96 (1.9, 4.62)	0.599 (0.270)	1.82 (1.07, 3.09)
log(popdensity)	1.009 (0.078)	2.74 (2.35, 3.20)	0.417 (0.043)	1.52 (1.39, 1.65)	0.959 (0.057)	2.61 (2.33, 2.92)
partyRepublican	−¿ 0.101 (0.222)	0.90 (0.57, 1.43)	−¿ 0.323 (0.122)	0.72 (0.57, 0.92)	1.093 (0.173)	2.98 (2.13, 4.19)
regionB	−¿ 1.879 (0.464)	0.15 (0.06, 0.38)	−¿ 0.509 (0.254)	0.60 (0.30, 1.20)	−¿ 1.108 (0.295)	0.33 (0.15, 0.72)
regionC	−¿ 2.621 (0.496)	0.07 (0.03, 0.19)	−¿ 1.673 (0.427)	0.19 (0.08, 0.44)	0.502 (0.376)	1.65 (0.79, 3.45)
regionD	−¿ 1.717 (0.527)	0.18 (0.06, 0.51)	−¿ 0.574 (0.269)	0.56 (0.27, 1.16)	0.242 (0.401)	1.27 (0.58, 2.80)
regionE	−¿ 1.941 (0.461)	0.14 (0.06, 0.35)	0.884 (0.324)	2.42 (1.28, 4.57)	−¿ 1.925 (0.402)	0.15 (0.07, 0.32)
regionF	−¿ 1.520 (0.522)	0.22 (0.08, 0.61)	0.629 (0.367)	1.88 (0.91, 3.85)	0.814 (0.444)	2.26 (0.95, 5.39)
regionG	−¿ 2.886 (0.647)	0.06 (0.02, 0.20)	0.363 (0.361)	1.44 (0.71, 2.92)	−¿ 1.536 (0.444)	0.22 (0.09, 0.51)
regionH	−¿ 2.221 (0.681)	0.11 (0.03, 0.41)	0.374 (0.396)	1.45 (0.67, 3.16)	−¿ 1.329 (0.570)	0.26 (0.09, 0.81)
regionI	−¿ 3.509 (1.117)	0.03 (0.00, 0.27)	0.657 (0.479)	1.93 (0.75, 4.93)	2.139 (0.476)	8.49 (3.34, 21.58)
regionJ	−¿ 2.527 (0.666)	0.08 (0.02, 0.29)	0.228 (0.396)	1.26 (0.58, 2.73)	−¿ 0.213 (0.490)	0.81 (0.32, 2.07)
GovResponse	−¿ 0.028 (0.018)	0.97 (0.94, 1.01)	−¿ 0.030 (0.009)	0.97 (0.95, 0.99)	−¿ 0.020 (0.013)	0.98 (0.96, 1.00)
Constant	−¿ 5.171 (1.202)	0.01 (0.00, 0.07)	−¿ 1.308 0.684	0.35 (1.35, 0.09)	−¿ 5.115 (0.934)	0.01 (0.00, 0.04)

We find that the clusters can be roughly described as follows:

- **C1:** low death rates throughout much of the pandemic; found mostly in Upper Midwest and mountain states,
- **C2:** high death rates in spring 2020 with another spike in December 2020/January 2021; found mostly in the northeast and other large cities,
- **C3:** low death rates until fall 2020, followed by peak in December 2020; spread throughout the U.S. with concentrations in Central Midwest and Great Lakes, and
- **C4:** steady death rates from late summer through December 2020, followed by a peak in January; spread throughout the U.S. with concentrations in California, the Southwest and Southeast.

The SVI Theme 3: Minority Status & Language is significantly associated with clustering in C2 vs. C1, yielding an odds ratio of 38.90. Counties with high levels of SVI Theme 3 are strongly associated with membership in C2 compared to C1. All CDC regions (B through J) showed a significant, negative association with C2 vs. C1, indicating that being located outside of region A (the Northeast, baseline category for region) is associated with a lower odds of clustering in C2 vs. C1. This is consistent with our initial finding from the map in Figure 4 which showed that the counties in C2 were primarily located in the Northeast.

The variable with the strongest positive association to C3, relative to C1, is SVI Theme 4: Housing & Transportation. Population density is also significant and positively related to C3. The governor's party is significant and negatively associated with C3, indicating that counties in states with Republican governors are associated with lower odds of clustering in C3 than in C1. The government response is also significant and negatively related to membership in C3, but the effect is small. Among the regions, the coefficient for region C (North Carolina, South Carolina, Georgia, and Florida) is significant and negative; thus, counties in these states are associated with lower odds of classifying in C3 than in C1. On the other hand, Region E is significant and positive which suggests counties in Minnesota, Wisconsin, Illinois, Indiana, Michigan, and Ohio are associated with a higher odds of clustering in C3.

SVI Theme 1 is not significant for membership in any of clusters C2 through C4; however, three of the SVIs (Household Composition & Disability, Minority Status & Language, and Housing & Transportation) are significant and positively associated with membership in C4. In addition, counties located in states with Republican governors are also associated with a higher odds of classification in C4 relative to C1. Among the CDC regions, regions I (California, Nevada, and Arizona) and F (New Mexico, Texas, Oklahoma, and Louisiana) have positive coefficients. Regions B, E, G, and H have significantly negative coefficients. The logarithm of population density is also a significant predictor for classification in C2, C3, and C4, relative to C1, which indicates that low population density is associated with clustering in C1.

Overall, the multinomial regression model correctly classified 61.25% of the counties into one of the four clusters. Table 4 gives the in-sample predictive performance of the multinomial regression model broken down by cluster. The balanced accuracy is similar for all four clusters, ranging from 0.63 to 0.80. A more nuanced view of the performance can be seen from the sensitivity and specificity. The model performs well in correctly classifying counties in Cluster C4 (Sensitivity = 0.74), which shows a sustained emergence in deaths beginning in late summer 2020. The model also performs well in classifying counties in Cluster C1 (Sensitivity = 0.71), counties with few deaths. However, it has only moderate ability to correctly classify counties into Clusters C2 and C3 (Sensitivity = 0.42 and 0.39, respectively). Note that the sensitivity performance for clusters C2 and C3 exceeds the expected sensitivity of 0.25 that would be obtained from random allocation among four classes in a balanced or imbalanced multiclass classification problem (see [21] for more details). In terms of specificity, the model performs well at identifying which counties are not in clusters C1 −¿ C4 with specificity values ranging from 0.71 −¿ 0.98. Figure 6 shows the distribution of the accuracy of the multinomial logistic model in predicting cluster membership. Counties that are correctly predicted from the model are indicated in a light color, while those that are incorrectly predicted are indicated in a dark color. The model provides some insight into the patterns across the US, but additional data are needed to more accurately classify counties in terms of the pattern of death rates due to COVID-19.

Table 4. The predictive performance of the multinomial regression model for each cluster.

	Balanced Accuracy	Sensitivity	Specificity

Cluster 1	0.71	0.71	0.71
Cluster 2	0.70	0.42	0.98
Cluster 3	0.63	0.39	0.88
Cluster 4	0.80	0.74	0.86

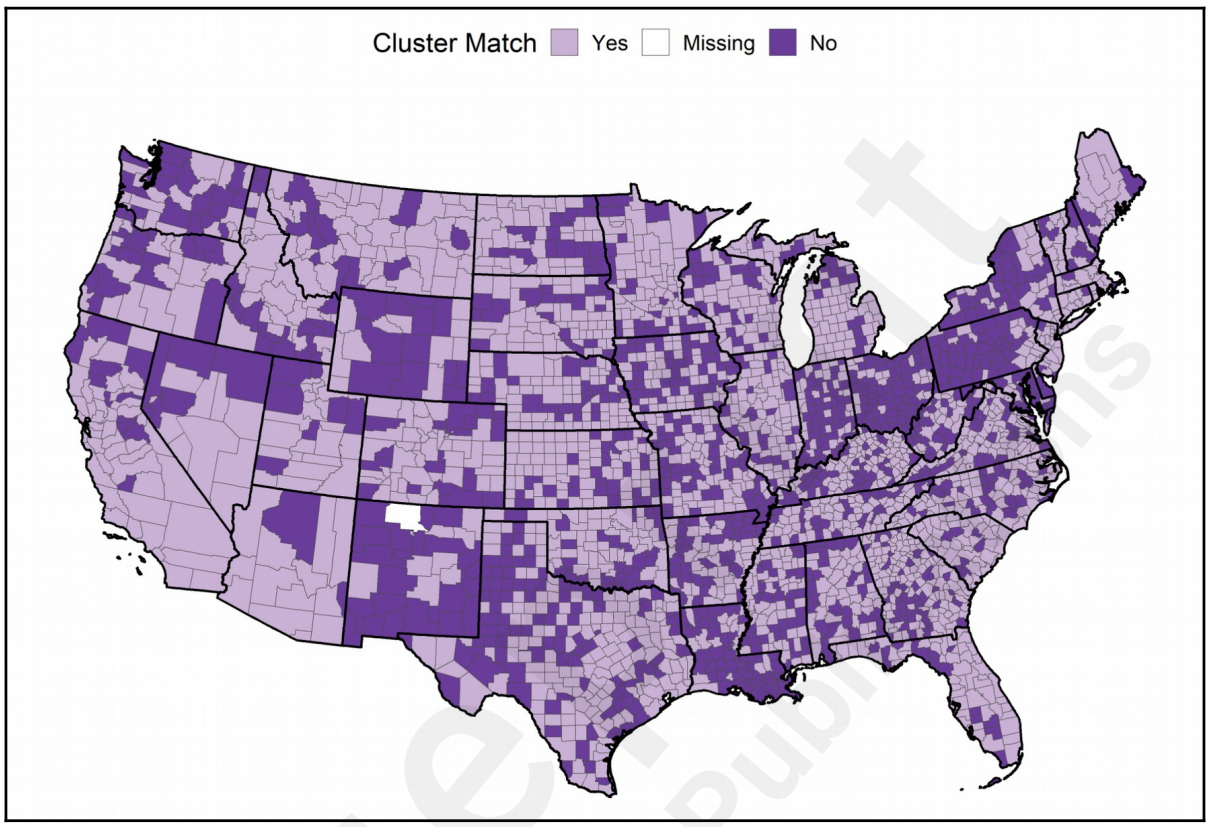


Figure 6. Map of the prediction accuracy of the multinomial logistic model describing the time series cluster solution. Counties in a light color (labeled Yes) were correctly classified with the model. Counties in a dark color (labeled No) were incorrectly classified. Rio Arriba County, New Mexico (in white) was not classified due to missing data. For an interactive version of this map, please see Section 4.2.4 in [20]..

Discussion

This research provides a framework for understanding the pattern of COVID-19 related deaths across the US. Using time series clustering with county-level data on the occurrence of COVID-19 related deaths, we observed four distinct patterns from March 1, 2020 to February 27, 2021. The second stage of our analysis revealed that these patterns can be partially explained by region as well as social and political predictors.

Our findings add to the literature on the relationship between COVID-19 outcomes and vulnerable populations [22–24]. The largest number of counties in the US experienced few deaths during the study period (cluster C1). These counties were, on average, at or below the median of all measures of social vulnerability. With lower population densities, and spread throughout the US, C1 counties served as our model baseline.

Cluster C3 (low death rates until fall 2020, peaking in December 2020) had the second largest number of counties. C3 counties are spread across much of the country, but have concentrations in the Great Lakes and Central Midwest regions. Interestingly, few incidences of C3 occur in the Southeastern U.S. and along the eastern seaboard from Washington DC to Massachusetts. Like C1,

counties in C3 had SVI measures below the median, on average. These counties experienced a single late wave in COVID-19 deaths beginning in late October 2020 that declined by the end of the study period. There were a few distinguishing features between counties classifying in C3 vs. C1: higher population density, Democratic state leadership, location outside of the Southeast, location in the Great Lakes region, and higher vulnerability in the SVI Housing & Transportation theme. This index indicates a higher incidence of multi-unit housing, mobile homes, crowding, lack of vehicle, or group living situations.

The 226 counties that clustered in C2 (high death rates in spring 2020 and December 2020/January 2021) were mostly in populous counties in the Northeast, Washington State, southeast Louisiana, including New Orleans, and the for corners region of Arizona and New Mexico. C2 counties experienced an early outbreak of deaths followed by a second wave beginning in November 2020 but few deaths in summer 2020. These counties showed a strong relationship with the SVI Minority & Language theme, indicating a large percentage of residents who are minority or non-native English speakers.

Cluster C4 (steady death rates beginning late summer, peaking in January) was located throughout the U.S. with concentrations in the Southeast and Southwest. The counties in C4 showed a steady incidence of deaths beginning in late summer 2020 that continued through the study period. C4 counties were, on average, above the median on all SVI themes, and three of the four themes were significant in classifying counties in C4 vs. C1. Specifically, the themes related to Household & Disability, Minority & Language, and Housing & Transportation all showed a positive association with this sustained pattern of COVID-19 related deaths. The majority (74%) of these counties were located in Republican led states.

The local patterns in COVID-19 related deaths suggest that local-level factors including geographic, demographic, and social vulnerability characteristics relate to adverse outcomes from COVID-19. There are several limitations to this research. These include the observational nature of the study which was conducted as the pandemic continues to emerge. The retrospective, secondary use of data makes it impossible to infer causation from our model. Outbreaks and adverse outcomes changed over time as local and national governments adopted new policies and vaccines to react to the emerging pandemic. Further, the Government Response Index is available only at the state level, and is constant across all counties within a state. Using a state-level predictor to explain cluster membership at the county level could lead to an ecological fallacy.

Despite these limitations, this exploratory study revealed new insights into the most severe outcome of the COVID-19 pandemic. The identification of four distinct patterns of death incidences in 3,108 U.S. counties provides evidence of the differences in the realization of severe outcomes from the pandemic. The U.S. is a demographically and politically diverse nation, and it is important to understand the differences in pandemic-related outcomes across communities. By examining the relationship between county-level predictors and membership in the four cluster patterns, we show that there are important demographic, political, and socioeconomic differences related to death patterns across the U.S.

Acknowledgments

Our data acquisition and computations were supported in part by the Ohio Supercomputer Center (grant # PZS1007).

Abbreviations

C1: Cluster 1
C2: Cluster 2
C3: Cluster 3
C4: Cluster 4

CDC: Center for Disease Control and Prevention
COVID-19: Coronavirus disease 2019
OR: odds ratio
SD: standard deviation
SVI: social vulnerability index

Supplementary Materials

The county-level COVID-19 deaths data is extracted using the COVID19 **R** package [1], which extracts the confirmed deaths from [25]. The cross-sectional dataset containing the predictors used in the multinomial regression has been compiled by the authors from disparate sources and is available at [26]. Version 4.0.4 of the **R** statistical software system was used for all processing and analysis of data. A reproducible workflow of our analysis is made available using **R** Markdown, and is hosted at [20], following the best practices of Jalali et al. [27] in reporting and documenting analyses for COVID-19.

References

1. Guidotti E, Ardia D. COVID-19 Data Hub. J Open Source Softw The Open Journal; 2020 Jul 10;5(51):2376.

2. Ahmed R, Williamson M, Hamid MA, Ashraf N. United States County-level COVID-19 Death Rates and Case Fatality Rates Vary by Region and Urban Status [Internet]. Healthcare. 2020. p. 330. [doi: 10.3390/healthcare8030330]

3. Gollwitzer A, Martel C, Brady WJ, Pärnamets P, Freedman IG, Knowles ED, Van Bavel JJ. Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic. Nat Hum Behav 2020 Nov;4(11):1186–1197. PMID:33139897

4. Baccini L, Brodeur A. Explaining Governors' Response to the COVID-19 Pandemic in the United States [Internet]. American Politics Research. 2021. p. 215–220. [doi: 10.1177/1532673x20973453]

5. Le NK, Le AV, Brooks JP, Khetpal S, Liauw D, Izurieta R, Ortiz MR. Impact of government-imposed social distancing measures on COVID-19 morbidity and mortality around the world [Internet]. [doi: 10.2471/blt.20.262659]

6. Megahed FM, Allison Jones-Farmer L, Rigdon SE. A retrospective cluster analysis of COVID-19 cases by county [Internet]. [doi: 10.1101/2020.11.12.379537]

7. Covid-19: The global crisis — in data [Internet]. [cited 2021 Jul 17]. Available from: https://ig.ft.com/coronavirus-global-data/

8. National Center for Chronic Disease Prevention and Health Promotion Regions [Internet]. Centers for Disease Control and Prevention (CDC) US Department of Health and Human Services. [cited 2020 Sep 19]. Available from: https://www.cdc.gov/coordinatedchronic/docs/nccdphp-regions-map.pdf9. Hale T, Atav T, Hallas L, Kira B, Phillips T, Petherick A, Pott A. Variation in US states responses to COVID-19. Blavatnik School of Government 2020;

10. Oxford COVID-19 Government Tracker Methodology [Internet]. 2020. Available from: https://github.com/OxCGRT/covid-policy-tracker/blob/master/documentation/index_methodology.md

11. Agency for Toxic Substances and Disease Registry. CDC Social Vulnerability Index [Internet]. US Department of Health & Human Services Centers for Disease Control and Prevention. 2020

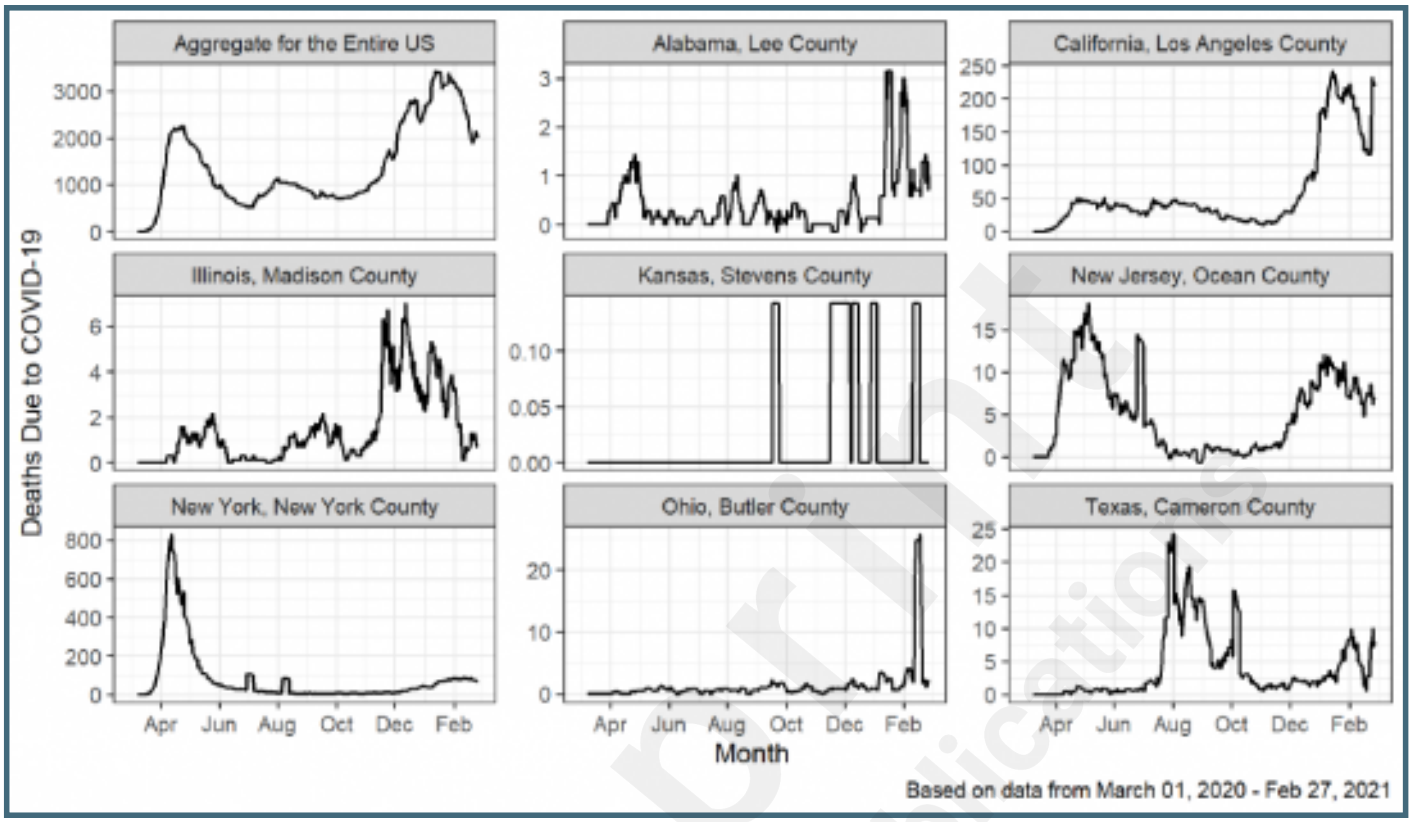
[cited 2021 Mar 21]. Available from: https://www.atsdr.cdc.gov/placeandhealth/svi/index.html

12. Flanagan BE, Gregory EW, Hallisey EJ. A social vulnerability index for disaster management. management [Internet] degruyter.com; 2011;8(1). Available from: https://www.degruyter.com/document/doi/10.2202/1547-7355.1792/html
13. Warren Liao T. Clustering of time series data—a survey. Pattern Recognition, Elsevier; 2005 Nov 1;38(11):1857–1874.
14. Bellman R. *Adaptive Control Processes: A Guided Tour*, Volume 2045 in the series Princeton Legacy Library, 2015. Available online from: https://doi.org/10.1515/9781400874668
15. Aghabozorgi S, Shirkhorshidi AS, Wah TY. Time-series clustering--a decade review. Inf Syst Elsevier; 2015;53:16–38.
16. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. Journal of Statistical Software, Articles 2014;61(6):1–36.
17. Charrad M, Ghazzali N, Boiteau V, Niknafs A. R Package `NbClust' 3.0 [Internet]. Available from: https://cran.r-project.org/web/packages/NbClust/NbClust.pdf
18. Hosmer Jr., D.W., Lemeshow, S., Sturdivant, R.X. *Applied Logistic Regression*, 3rd Ed. John Wiley & Sons, Hoboken, New Jersey, 2013.
19. Ripley B, Venables W. R package 'nnet' 7.3-15 [Internet]. Available from: https://cran.r-project.org/web/packages/nnet/nnet.pdf
20. Megahed, F.M., Jones-Farmer, A., Rigdon, S. A reproducible GitHub page for a two-stage modeling framework for analyzing COVID-19 deaths by county; 2021. Available from https://fmegahed.github.io/covid_deaths.html
21. Megahed, F.M., Jones-Farmer, A., Rigdon, S. A numerical study to examine the obtained sensitivity for arbitrary and proportional guessing scenarios for a 4-class classification problem; 2022. Available from https://fmegahed.github.io/research/sensitivity/simulation_sens_computation_multiclass_baselines.html
22. Chen JT, Krieger N. Revealing the Unequal Burden of COVID-19 by Income, Race/Ethnicity, and Household Crowding: US County Versus Zip Code Analyses. J Public Health Manag Pract Wolters Kluwer; 2021;27 Suppl 1, COVID-19 and Public Health: Looking Back, Moing Forward(1):S43–S56. PMID:32956299
23. Stokes AC, Lundberg DJ, Elo IT, Hempstead K, Bor J, Preston SH. COVID-19 and excess mortality in the United States: A county-level analysis. PLoS Med Public Library of Science San Francisco, CA USA; 2021 May;18(5):e1003571. PMID:34014945
24. Khanijahani A. Racial, ethnic, and socioeconomic disparities in confirmed COVID-19 cases and deaths in the United States: a county-level analysis as of November 2020. Ethn Health Taylor & Francis; 2021 Jan;26(1):22–35. PMID:33334160
25. JHU CSSE COVID-19 Data; 2021. A GitHub repository available from https://github.com/CSSEGISandData/COVID-19.
26. Megahed, F.M. Covid-19 Deaths; 2021. A GitHub repository available from https://github.com/fmegahed/covid19-deaths/tree/master/Data/Output
27. Jalali MS, DiGennaro C, Sridhar D. Transparency assessment of COVID-19 models. Lancet Glob Health Elsevier; 2020 Dec;8(12):e1459–e1460. PMID:33125915

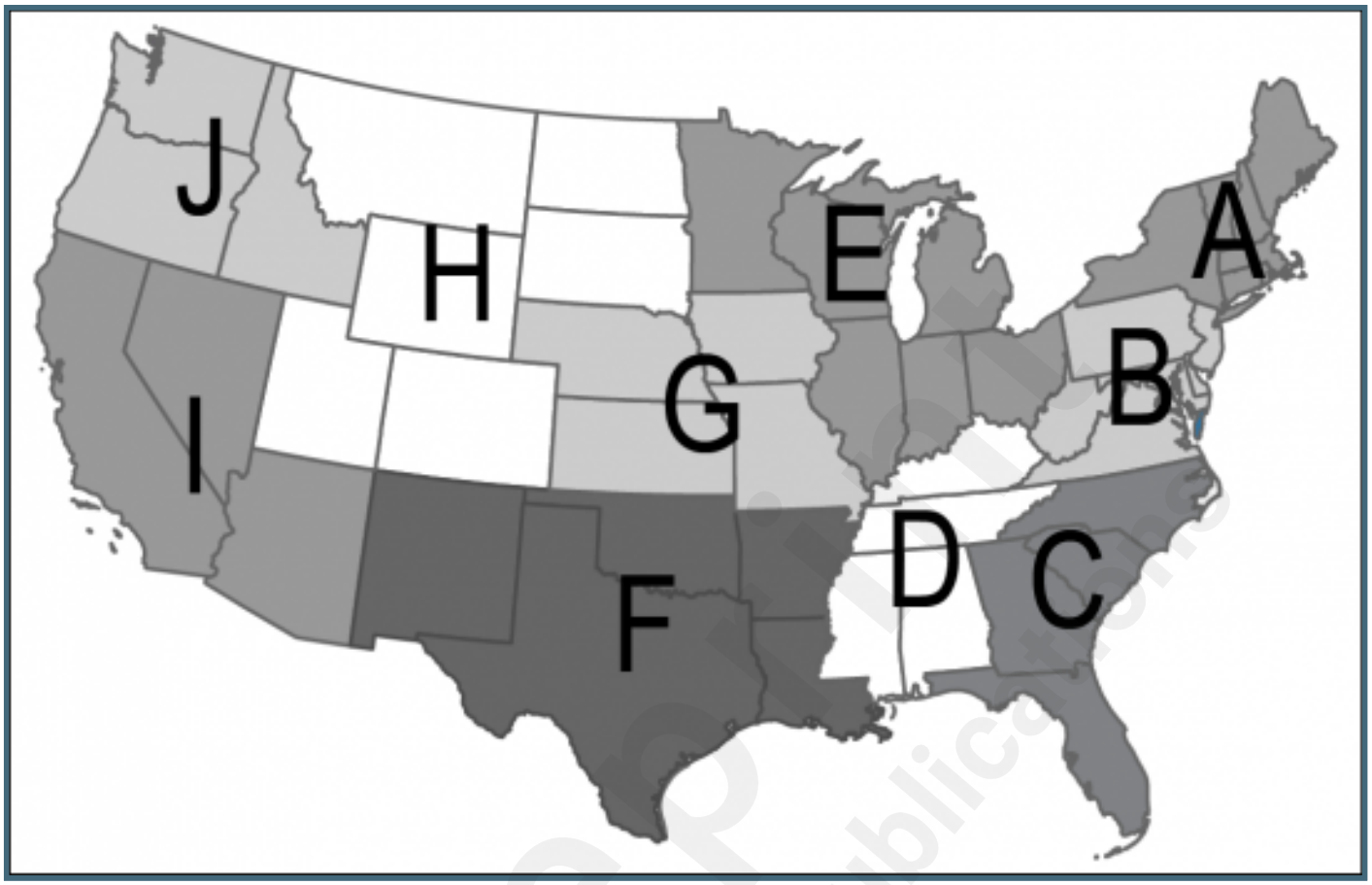
Supplementary Files

Figures

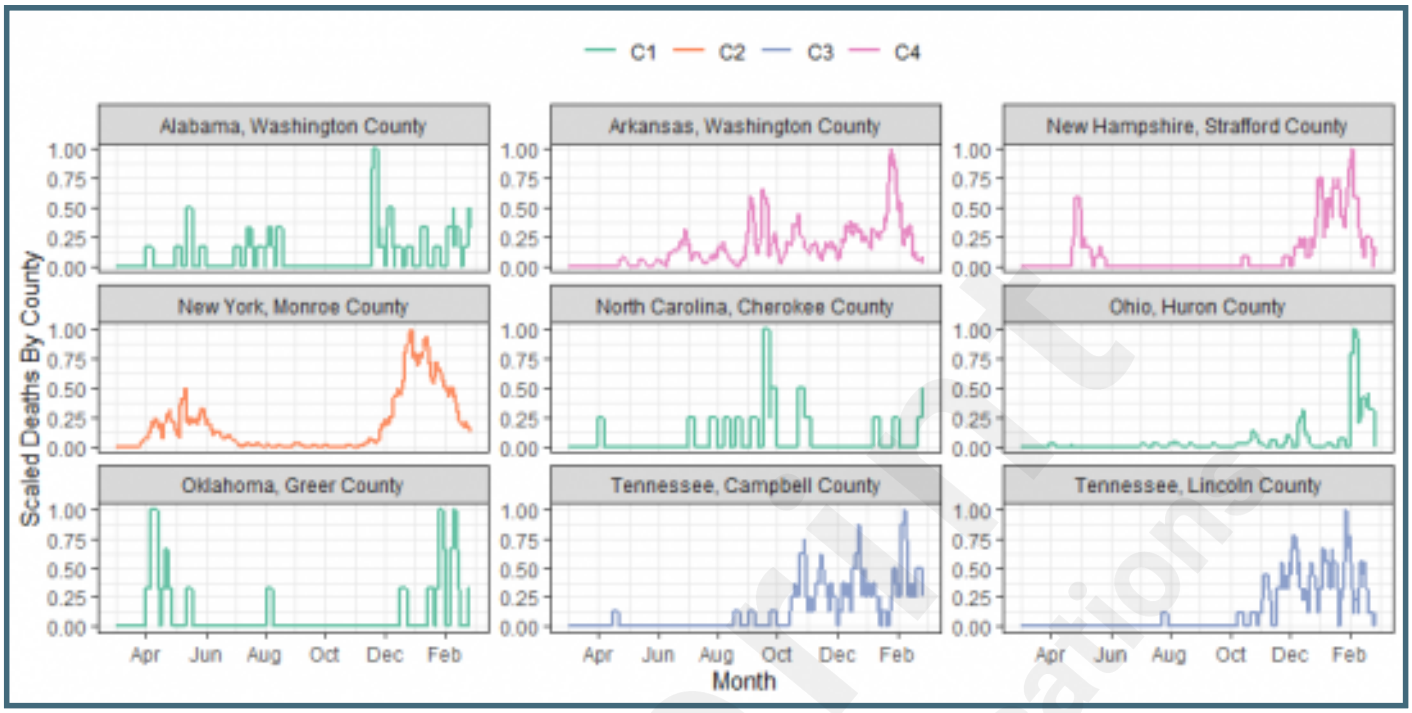
The time series profiles of the 7-day moving average of new COVID-19 deaths for the entire US and eight sample counties.



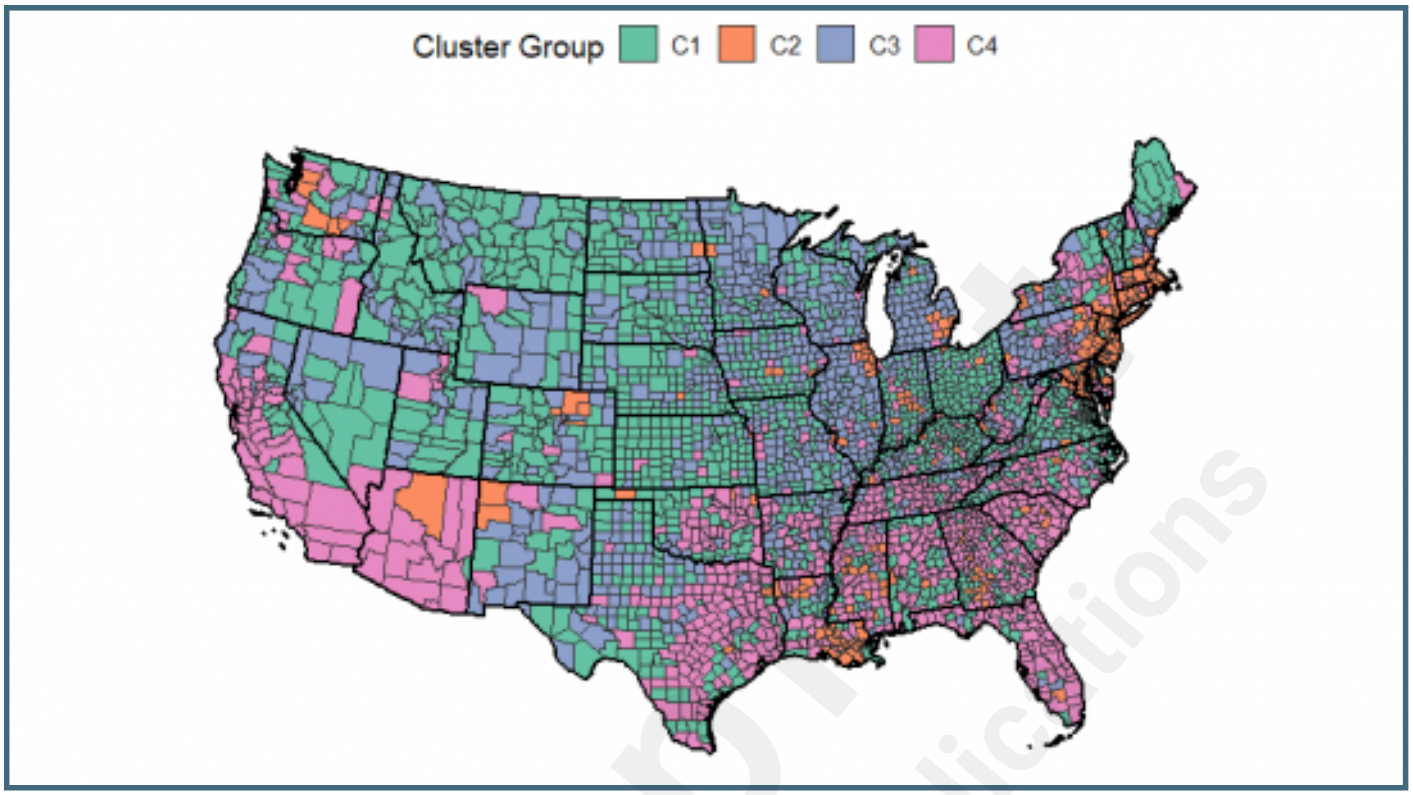
The ten CDC regions.



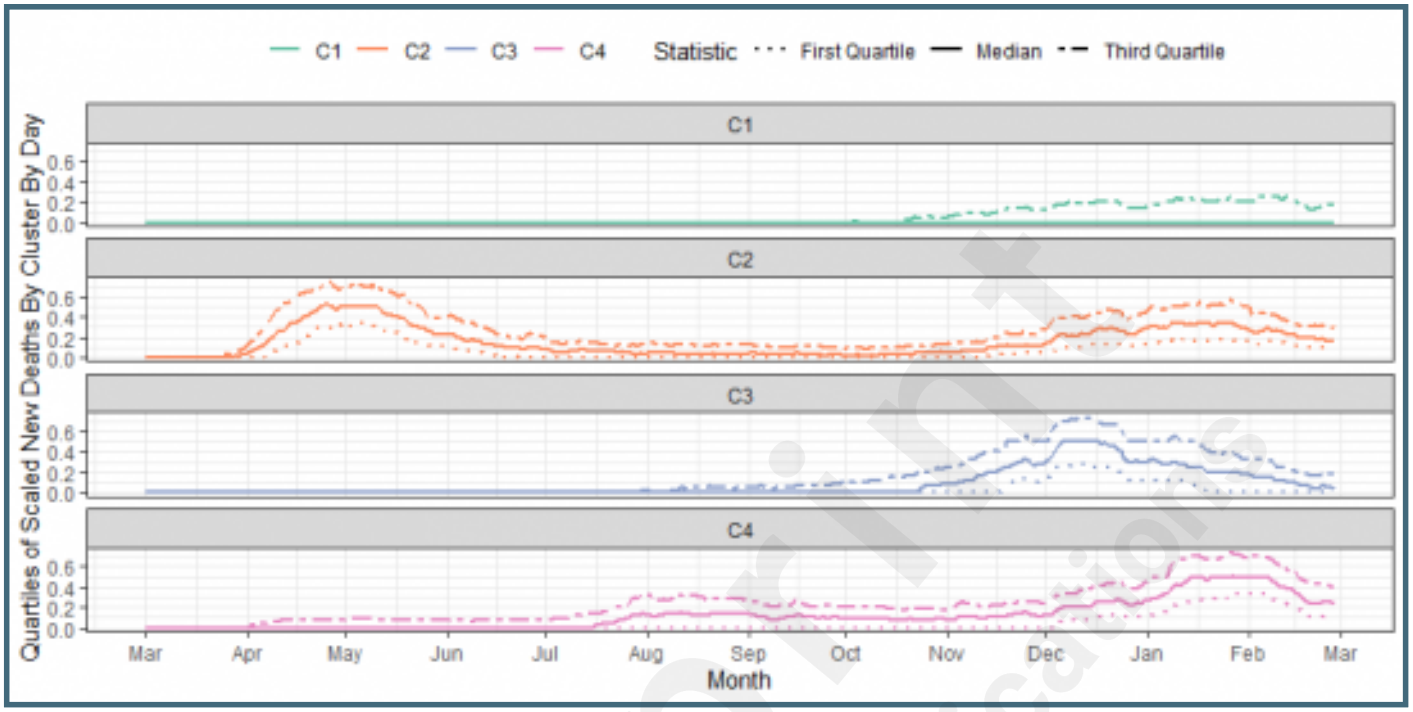
The time series profiles of the scaled 7-day moving average of new COVID-19 deaths for nine sample counties.



Map of the four scaled time-series profile clusters of COVID-19 deaths by county in the contiguous United States. For an interactive color version of this map, please see Section 3.3.3 of [20].



A summary plot, where the median scaled time-series profile for each cluster is depicted using the ‘solid’ bold line. The first and third quartiles are shown through ‘dotted’ and ‘two-dash’ lines, respectively.



Map of the prediction accuracy of the multinomial logistic model describing the time series cluster solution. Counties in a light color (labeled Yes) were correctly classified with the model. Counties in a dark color (labeled No) were incorrectly classified. Rio Arriba County, New Mexico (in white) was not classified due to missing data. For an interactive version of this map, please see Section 4.2.4 in [20].