Title

Characterizing pandemic waves: A latent class analysis of COVID-19 spread across US counties.

Authors

Sarwar Uddin, Md Yusuf
Rafiq, Rezwana

Publication Date
2022-10-01

DOI
10.1016/j.patrec.2022.08.017

Peer reviewed

Pattern Recognition Letters 162 (2022) 31–39

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Characterizing pandemic waves: A latent class analysis of COVID-19 spread across US counties

Md Yusuf Sarwar Uddin[a,*], Rezwana Rafiq[b]

[a] University of Missouri-Kansas City, Kansas City, MO 64110, United States
[b] University of California, Irvine, CA 92617, United States

ARTICLE INFO

Article history:
Received 21 April 2021
Revised 21 June 2022
Accepted 29 August 2022
Available online 31 August 2022

Edited by: Maria De Marsico

Keywords:
COVID-19 pandemic
Pandemic waves
Spread patterns
Latent class analysis
Human mobility

ABSTRACT

The spread of the COVID-19 pandemic is observed to follow the shape of “waves” (i.e., the rise and fall of population-adjusted daily new infection cases with time). Different geographic regions of the world have experienced different position and span of these waves over time. The presence and strength of these waves broadly characterize the dynamics of the pandemic spread in a given area, so their characterization is important to draw meaningful intervention and mitigation plans tailored for that area. In this paper, we propose a novel technique to represent the trend of COVID-19 spread as a sequence of a fixed-length text string defined on three symbols: R (rise), S (Steady), and F (fall). These strings, termed as *trend strings*, enabled us searching for specific patterns in them (such as for waves). After analyzing county-level infection data, we observe that, US counties—despite their wide variation in trend strings—can be grouped into a number of heterogeneous classes each of which might have a representative COVID spread pattern over time (in terms of presence and propensity of waves). To this end, we conduct a latent class analysis to cluster 3142 US counties into four distinct classes based on their wave characteristics for one year pandemic data (January 2020 to January 2021). We observe that counties in each class have distinct socio-demographics, location, and human mobility characteristics. In short summary, counties have differing number of waves (class 1 counties have only one wave and class 3 counties have three) and their positions also vary (class 1 had the wave later in the year whereas class 3 had waves throughout the year). We believe that this way of characterizing pandemic waves would provide better insights in understanding the complex dynamics of COVID-19 spread and its future evolution, and would, therefore, help in taking class-specific policy interventions.

1. Introduction

The United States is ravaged by the COVID-19 pandemic. Since the first case recorded on January 22, 2020, the country has reported over 86 millions confirmed cases with a death toll exceeding one million [1]. At the onset of more than two full years into the pandemic, it is important to understand the dynamics of the pandemic spread as it have evolved since January 2020. The literature in epidemiology reports that epidemics of infectious diseases are frequently characterized by the presence of one or more “waves” [2–4], where a “wave” represents the rise of the daily infection cases up until a high peak and then declining over time to reach a bottom peak (followed by rising again). The understanding of when and how these waves emerge is very critical to take effective intervention and control measures. Interestingly, different geographic areas of the world (e.g., countries, states, or counties) have experienced different positions and spans of these waves over time as so in the United States, which we study in this paper.

Related work: The rise and fall of cases have been well documented in pandemic history. The famous 1918 Spanish influenza pandemic (1918–1920) reportedly spread through several US and European cities in multiple waves with varying frequency and timing of individual epidemic peaks [5–7]. Mishra et al. [2] recollect pandemic waves in history (from Bubonic plague in 14th century, to Smallpox to present day COVID-19) and analyzed disease spread characteristics.

Various modeling efforts, both mathematical and empirical, have been reported in the literature that attempted to model and describe the dynamic evolution of pandemic spread in the US and other countries. Studies include agent-based simulations and a multiple wave model [8], data-driven model based on Susceptible-Infectious-Removed (SIR) model applied at US county-level and its

[*] Corresponding author.
E-mail address: muddin@umkc.edu (M.Y. Sarwar Uddin).

https://doi.org/10.1016/j.patrec.2022.08.017

other variants such as deterministic SEIR (SIR plus Exposed compartment) model for US state-level infections [9,10], SIR simulations [11], stochastic transmission model [12], spatiotemporal “risk source” model [13], spatiotemporal multivariate time series model (US state-level) [14], Bayesian hierarchical spatiotemporal model (US county-level) [15], and detection and forecast of pandemic waves [16]. Articles [17,17] compared pandemic waves in 2020 and 2021 in different countries and regions and demonstrated that waves differ significantly across regions. Focused studies for specific regions are also available: Africa [18], South-East Asia [19], and Europe [20,21]. The effects of COVID-19 pandemic waves on different socio-economic sectors are also analyzed in the literature: mental health [22], transportation [23], and economy and government responses [24].

While these efforts attempted to model the progress and dynamics of COVID spread for a given geographical unit (country, state, or county), only a few prior works tried to *group* geographic regions based on their COVID spread patterns. A recent study [14] reported the heterogeneity of COVID-19 transmission patterns across US states in association with the weather or climate factors from March to September 2020. Based on the pandemic data from 18 countries, [8] suggested that the COVID-19 pandemic can be successfully modeled as a series of epidemic waves and identified the presence of three such waves in those countries. Rios et al. [25] applied hierarchical clustering on world countries based on the number of new cases and deaths and tracked the temporal transition of countries across clusters as the COVID-19 pandemic evolved. In contrast, we captured the temporal variation of infection cases across US counties as a *pattern string* (discussed later) and then clustered the counties based on those pattern string characteristics.

The definition of pandemic waves vary in literature. In addition to usual rise and fall of daily new cases as adopted by most articles [2,17,18,21], other formative definitions exist. Zhang et al. [26] defined waves in terms of a ratio R (the average number of people infected by a single infectious individual) where $R > 1$ and $R < 1$ indicate up-trend and down-trend, respectively. Fisayo and Tsukagoshi [27] designates waves as *eventual* effect of on human life: direct effect leading to deaths and disability (first wave), effects due to measures taken to limit the spread (second wave), and social determinants of health, and its effects on the next generation (third wave).

Our contribution: In this paper, we analyze the heterogeneity of COVID-19 spread patterns across US counties and their association with particular socio-demographic, location, and human mobility characteristics of those counties over the one year of the pandemic from January 2020 to January 2021. One of the challenges to analyze these spread patterns is that these patterns vary widely across US counties. We postulate that US counties, despite the variations, can be grouped into a number of heterogeneous classes so that same class counties share similar spread patterns and socio-demographic, location, and human mobility characteristics but the characteristics across classes vary. Towards this end, we made two contributions. First, we propose a novel technique to represent the pandemic spread in a county as a *text string* of a fixed length of three symbols: R (rise), F (fall), and S (Steady), which indicates the year around increase and decline of daily new infection cases in that county. This representation enables us to locate the presence of pandemic waves and to determine their spans and heights in the respective counties. Second, we conduct a clustering analysis using LCA (Latent Class Analysis) to group counties into classes based on their pandemic spread patterns and the associated socio-demographic, location, and mobility factors. Unlike other popular clustering techniques, such as hierarchical clustering [25] and K-means clustering [28], LCA not only produces clusters, but also provides probabilistic support for cluster membership in terms of designated attributes [29], which is very important for analyzing and explaining clusters and their formations.

We conduct the clustering at the county level because counties are arguably the smallest administrative units in the US for coordinating COVID responses. We believe that this way of clustering would help to build different prediction models for different groups for better pandemic forecast and thus would help in taking effective county-level localized preventive measures and resource allocation (e.g., counties with similar trends can choose a similar set of actions).

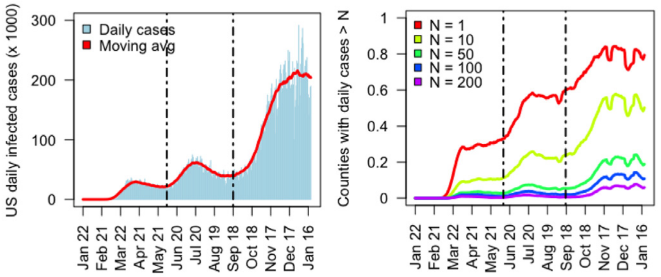
2. Material and methods

2.1. Data sources and time frame

The datasets used for this study are drawn from heterogeneous sources as listed in Table 1. The sources include John Hopkins COVID-19 data repository [30], Maryland Transportation Institute (MTI) COVID-19 Impact Analysis Platform [31], Google COVID-19 Community Mobility reports [32], and US Census Bureau 2018 [33]. The John Hopkins US dataset contains the county-level daily infection data (daily cumulative confirmed, deaths, and recovered) since the first case in the US (Jan 22, 2020). For the MTI dataset, we extracted data from their publicly available web portal. Google COVID-19 Community Mobility reports contain the percentage change in visitor locations from a baseline in different geographic areas of the world including the US (5‘week period from Jan 3 to Feb 6, 2020 was used as the baseline). The report categorizes activity places into a set of commonly referred types, such as retail and recreation, groceries and pharmacies, parks, transit stations, workplaces, and residential. Furthermore, the socio-economic and location attributes of counties are obtained from the US Census Bureau 2010. Multiple datasets are joined using county FIPS (Federal Information Processing Standard) codes.

Table 1
US County-level Summary Statistics (Number of counties, $N = 3142$).

Variable	Source	Description
COVID-19 Infection rate	[30]	Number of daily new cases per 1000 people
COVID-19 status		
Test completed	[31]	Number of COVID-19 tests completed per 1000 people
Imported cases	[31]	Number of daily external trips by infectious persons from out of county
Human Mobility		
Social distancing index (An integer from 0–100 measuring the extent of social distancing)	[31]	Weighted sum of % of people staying home and reduction of human movement in different aspects, such as work trips, non-work trips, and distance traveled. ‘0’ indicates no social distancing; ‘100’ denotes all residents are staying at home and no visitors are entering the county
Percentage change in non-workplace visits	[32]	Percentage change in visits to non-workplace (retail, recreation, grocery and pharmacy) with respect to the baseline (Jan 3 - Feb 6, 2020)
Socio-demographic and Location Characteristics		
Age above 60	[33]	% of people older than 60 years
African-Americans	[33]	% of African-Americans
Population Density	[33]	Population density as total number of people/total land area (sqmile)
Metro status	[34]	Metro status based on 2013 NCHS urban-rural classification; 1=metro, 0=nonmetro



(a) US daily cases (b) Fraction of counties

Fig. 1. (a) Daily infection cases over the entire US, (b) Fraction of counties exceeding a certain infection count.

Figure 1 a shows the daily *new* infection cases reported in the entire US for one full year from January 22, 2020 to January 21, 2021. As we can observe, there have been a series of ups and downs in the daily infection cases suggesting that the pandemic passed through a series of "waves" (also known as sub-epidemics [8]). Figure 1a clearly demonstrates the presence of three such waves during the one year timeline: the first wave appears between January and May (Jan 22, 2020–May 30, 2020), the second one appears somewhat over the summer (Jun 1, 2020–Sep 18, 2020), and the third one spans the last four months of the year; Sep 19, 2020–Jan 21, 2021). We denote these three time spans as *early*, *mid*, and *late* time window of pandemic waves (we also use the term Window 1, Window 2, and Window 3 interchangeably). Moreover, Fig. 1b shows the fraction of US counties that have exceeded a certain number of daily new cases over the span of the year. It turns out that this fraction also grew in wave-like fashion with occasional jumps here and there as time progressed.

Although the overall US trend line shows a distinct pandemic waves, the county-level infection data are more irregular and curvy and, therefore, do not necessarily resort to clearly visible wave-like pandemic spread trends. We need to process the infection data in a certain way to extract the hidden wave-like spread pattern in them (which we describe next).

3. Method

In this study, we first develop a technique to characterize the COVID-19 spread across US counties based on the presence of wave patterns of daily infection cases over one year time span. Then, we cluster the counties based on their wave patterns.

3.1. Formation of pandemic wave patterns

The pandemic wave of a given geographic area corresponds to the rise/increase of daily new infection cases followed by the fall/decline in daily new cases within a certain span of time. To identify the presence of waves in country-level infection time-series data, we split the entire timeline (Jan 22, 2020 to Jan 21, 2021) into 12 equal length intervals (each interval of length around 30 days). For each interval, we determine whether the daily new infection cases per 1000 population in that interval are growing (*rise*), or declining (*fall*), or remaining as *steady*, by denoting them by three distinct symbols, R, F, and S, respectively. In that, we obtain a 12-letter textual string for each county that represents the year-long pandemic progress in that county. These text strings are then processed to locate the presence of one or more pandemic waves (if any) in respective counties. The symbols used in this paper are shown in Table 2:

Table 2
Notations used in this paper.

	Defining Pandemic Waves
i	Index to denote a county
N	Number of counties
ℓ	Time interval index
$I_d^{(i)}(t)$	Daily infection rate at time t at county i
$I_m^{(i)}(t)$	Moving average of daily infections
$\beta_\ell^{(i)}$	The slope of the trend line
τ	The cutoff threshold of slopes for RISE and FALL
$\Gamma^{(i)}$	Pattern string defined on alphabet {R, F, S}
	Latent Class Analysis
c	Index to denote a class/group (total C classes)
j	Attribute of a county
K_j	Possible values of attribute j
Y_{ijk}	Binary indicator if county i took value k for attribute j
X_i	Covariates of county i
ρ_c	Class probabilities or mixing probabilities
π_{cjk}	Class-conditional probabilities
η_c	Coefficients of covariates for class c

3.1.1. Trend lines of daily infections per county

We observe that county-level time-series data of the infection rates (daily new cases for 1000 population) are very irregular and zigzag in shapes as the reported cases vary widely from one day to the next day due to various reasons. In order to smooth out the variations and to filter out short-spanned instantaneous jumps in the time-series data, we take a *moving average* of the infection rates. Let $I^{(i)}(t)$ denote the time-series of daily new confirmed infections per 1000 population at day t for a county i, and $I_m^{(i)}(t)$ denote the moving average of the time-series values. The κ-day moving average is calculated as (for $\kappa = 2l + 1$, we use $\kappa = 15$):

$$I_m^{(i)}(t) = \frac{1}{\kappa} \sum_{\delta=-l}^{l} I_d^{(i)}(t + \delta) \tag{1}$$

The moving average of the infection rates along days constitutes the *trend line* of COVID growth in a given county. We then split the entire timeline (366 days) into a set of consecutive *intervals* each of which has an equal length and analyze the individual time segments for a possible *rise* and *fall* trend. Let the length of this interval be Δ days. In that, we obtain a series of intervals as 1, 2, 3, ..., L, in the entire timeline where D is the total number of days in the time-series data, and $L = \lceil \frac{D}{\Delta} \rceil$. We denote an arbitrary interval by ℓ and the associated range of days of the interval as $\mathcal{I}_\ell = (T_{\ell-1}, T_\ell]$, where $T_\ell = \ell \times \Delta$ and $T_0 = 0$. For one year of infections data at hand and with the consideration of 30-day interval ($\Delta = 30$), we obtain 12 such intervals. In that, the infection time-series data constitutes a sequence of 12 such consecutive intervals.

We subsequently find a piece-wise linear approximation of the curvy trend line per interval for a given county. For that, we run a linear regression on each interval of the time-series data to compute the slope of the line segment that best fits the moving average of the daily news cases with respect to the respective days in that interval. For a given interval ℓ, we fit the following linear equation for a county i:

$$I_m^{(i)}(t) = \alpha_\ell^{(i)} + \beta_\ell^{(i)} \times t, \;\; \text{for } T_{\ell-1} < t \le T_\ell \tag{2}$$

where $\alpha_\ell^{(i)}$ and $\beta_\ell^{(i)}$ are the intercept and slope of the estimated line, respectively, for interval ℓ for county i, which are estimated by the Ordinary Least Square (OLS) method.

3.1.2. String representation of trend lines

Once the slopes are obtained, we observe their values for the possible rise and fall trends in the infection rates. Ideally, a positive slope indicates a rise whereas a negative slope indicates a fall during the corresponding interval. The higher value of a positive slope (resp. lower value of a negative slope) means a higher degree

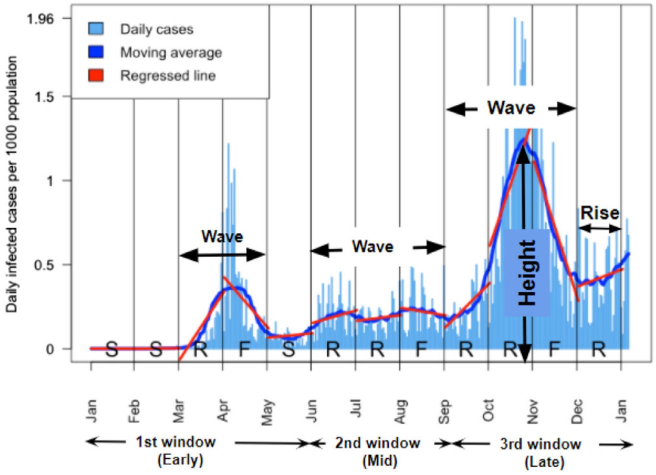


Fig. 2. Daily infection trend for Worcester county, Maryland (FIPS code 24047). The corresponding pattern string (SSR ⋯ FR) appears at the bottom.

of rise (resp. higher degree of fall). We also check the goodness-of-fit measure of the regression by observing its R-squared value (R^2), which is a value between 0 and 1, and a higher value indicates higher fit.

To this effect, we set a threshold τ, to designate a cutoff margin for positive and negative slopes to flag the trend as "rise" and "fall". In particular, if the positive slope is above τ and the R^2 value is higher than 0.75, we flag the change as "rise". In contrary, if the negative slope falls below $-\tau$ with $R^2 \geq 0.75$, we designate that as "fall". Otherwise, we designate the trend as "steady" indicating no significant upward or downward changes in daily infection cases during that interval. We use three symbols: R, F, and S, to denote the rise, fall, and steady respectively. In that, the trend of COVID-19 infection cases of each county can be represented as a *string* of length L defined on the alphabet {R, S, T}. We refer to this string as the *pattern string* and denote the pattern string as $\Gamma^{(i)}$ for county i. Each element of the pattern string, $\Gamma^{(i)}_\ell$, $\ell = \{1, 2, \dots, L\}$ is calculated as:

$$\Gamma^{(i)}_\ell = \begin{cases} \mathtt{R}, \text{ if } \beta^{(i)}_\ell > \tau \text{ and } R^2 \geq 0.75 \\ \mathtt{F}, \text{ if } \beta^{(i)}_\ell < -\tau \text{ and } R^2 \geq 0.75 \\ \mathtt{S}, \text{ otherwise} \end{cases} \tag{3}$$

The pattern string in a given county ultimately represents the time varying progress of COVID spread in that county. Figure 2 shows such a pattern string, SSRFSRRFRRFR, for a sample US county, Worcester, Maryland.

3.1.3. Identification of pandemic waves from pattern strings

The construction of the pattern strings from numerical infection rates enables us to analyze the pandemic trend in a very significant way. For instance, we can search for one of more sub-strings of specific patterns of our interests. In particular, we search for the occurrences of a "wave pattern" expressed by the following regular expression: R+(S*)F+. The pattern intends to match a sub-string in a given pattern string that contains a series of rises (R's), followed by zero or more steady cases (S's), and subsequently followed by one of more falls (F's). Each occurrence of this wave pattern in the pattern string corresponds to the presence of one pandemic wave in the corresponding county and multiple occurrences indicate multiple waves.

Obviously, each wave should have a rise followed by a fall. But, there are counties where only rises are observed without any fall suggesting that the daily infection cases are monotonically rising. This is usually observed during the late pandemic window (Window 3). We refer to this kind of trend as pure RISE, which is denoted by this regular expression: (R+(S)*)+ (a series of rises followed by steady cases but NO falls till the end). For Worcester county, as in Fig. 2, we observe that the county experienced three pandemic waves with different degree of intensities (i.e., heights) and a RISE at the end.

We observe that out of 3142 US counties, 1389 counties have one pandemic wave, whereas 1275, 265, and 4 counties have 2, 3, and 4 waves respectively (209 counties have zero waves). For each county, we determine the position, span, and height (the peak infection cases during a wave) of those pandemic waves. We then use these information to cluster the counties into a number of groups that we describe next.

3.2. Clustering of pandemic wave patterns

We conducted a Latent Class Analysis (LCA) [35,36] to partition 3142 US counties into a small number of heterogeneous groups or classes so that the counties belonging to the same class show similar characteristics in their pandemic waves, while maximizing the heterogeneity across groups. LCA *probabilistically* assigns counties to different classes based a set of categorical "indicator" variables that are to measure the differences and similarities among counties based on their pandemic spread pattern characteristics. LCA also takes a set of "covariates" to characterize the membership of different counties into different classes so as to analyze what socio-demographic, location, and human mobility characteristics of a county makes the county to belong to a certain class versus the others. LCA constructs two sub-models for the above two tasks and *jointly estimate* the model parameters. In the following we describe the LCA method followed by the application of LCA for the classification of counties and the fit statistics.

3.2.1. Latent class analysis

LCA is a mixture model that hypothesizes that there is an underlying *unobserved* categorical variable that divides a collection of entities (in our case, the US counties) into mutually exclusive and exhaustive latent classes [36,37]. Suppose each member of a population (here counties, indexed by i), contains J indicator variables (indexed by j), each of which can take a value from a set of K_j possible outcomes. Let Y_{ijk} a binary indicator, as such $Y_{ijk} = 1$ if county i takes kth outcome for its jth categorical variable, and $Y_{ijk} = 0$ otherwise. For a given number of classes, say C, LCA attempts to simultaneously compute: (a) the probability that a respondent falls into a certain class, denoted by ρ_c, for $c = 1, 2, \dots, C$, and (b) the class-conditional probability, denoted by π_{cjk}, that observation in class c produces the kth outcome on the jth variable. The probability of observing a certain respondent is, therefore, given by:

$$Pr(Y_i|\pi, \rho) = \sum_{c=1}^{C} \rho_c \prod_{j=1}^{J} \prod_{k=1}^{K_j} (\pi_{cjk})^{Y_{ijk}} \tag{4}$$

The parameters that the LCA model estimates are: (a) the class membership probability vector, $[\rho_c]$ (size C), and (b) the class-conditional probability matrix, $[\pi_{cjk}]$ (size $C \times \sum_j (K_j - 1)$). The parameters are computed via the maximum log-likelihood estimation (MLE). More formally, MLE tries to estimate ρ_c and π_{cjk} so as to maximize the likelihood of obtaining the observed Y_i's as they are, which is given by the following expression:

$$\mathcal{L} = \prod_{i=1}^{N} Pr(Y_i|\pi, \rho) \Rightarrow \ln \mathcal{L} = \sum_{i=1}^{N} \ln Pr(Y_i|\pi, \rho) \tag{5}$$

Putting Eq. (4) in the above, we obtain:

$$\ln \mathcal{L} = \sum_{i=1}^{N} \ln \left(\sum_{c=1}^{C} \rho_c \prod_{j=1}^{J} \prod_{k=1}^{K_j} (\pi_{cjk})^{Y_{ijk}} \right) \tag{6}$$

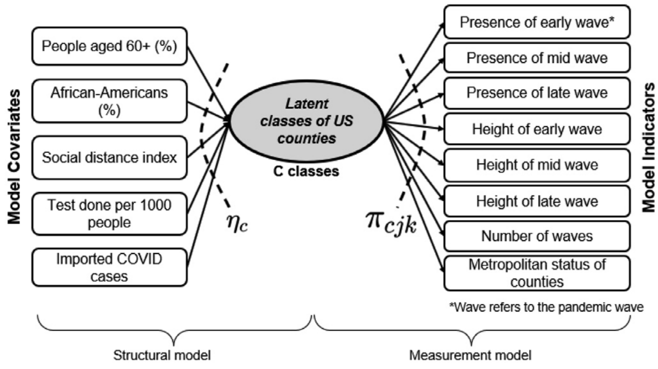
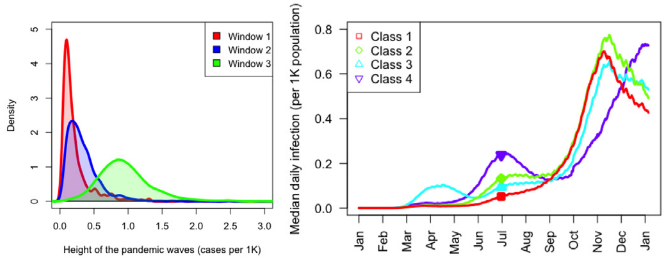


Fig. 3. Conceptual diagram of the Latent Class Analysis.



(a) Wave height

(b) Median infection cases

Fig. 4. Median daily new cases per class and density plot of wave heights.

LCA attempts to estimate ρ_c and π_{cjk} as to maximize the above. The class probabilities, ρ_c's, are however regressed from a set of "covariates" by applying the multinomial logistic (MNL) regression. In that, LCA finds a set of per class coefficient vectors, denoted as η_c of size $V+1$ for class c, instead of scalar ρ_c's. Therefore, the number of parameters that LCA jointly estimates is equal to $C\sum_j(K_j-1)+(V+1)(C-1)$, the sizes of the matrices $[\pi]$ and $[\eta]$, respectively (for detail, refer to [37]).

3.2.2. Model indicators and covariates

The latent class model developed in this study include both indicator variables and covariates. Indicator variables are used to define latent classes by estimating a measurement model whereas covariates are used to predict the probability of an county belonging to a latent class through the estimation of a structural model. Fig. 3 shows the conceptual latent class model in the combination of measurement and structural components with the set of indicator variables and covariates selected for this analysis. To capture the heterogeneity in COVID-19 pandemic spread patterns across different counties, we used various wave related attributes, such as the presence of pandemic waves in Window 1, Window 2, and Window 3, the total number of waves, the height of the waves by three categories: low, mid, and high based on whether their values fall below 33-percentile, between 33- and 66-percentile, and above 66-percentile value, respectively, among all the wave height values in the respective windows (Fig. 4a shows the height density in three windows). If more than one wave fall into one window, we consider the highest height among them. Note that only rise exists in Window 3. We also consider the metropolitan status of a county as one of the binary indicators. To understand the class membership profile we consider several active and inactive covariates that consist of various socio-demographic, location, and human mobility characteristics. In particular, these covariates include fraction of people aged above 60 years and fraction of African-Americans living in a county, population density, number of COVID tests completed, social distance index, number of inter-regional trips by infectious persons (imported COVID cases), cumulative infection cases, and changes in work and non-work place visits from baseline period.

3.2.3. Model estimation and fit statistics

We used poLCA (Polytomous Variable Latent Class Analysis) package [37] in R to estimate the LCA model parameters for our study. The package provides a set of goodness of fit measures for a fitted model, such as χ^2, AIC (Akaike Information Criterion) [38] and BIC (Bayesian Information Criterion) [37,39], which try to balance between over- and under-fitting of the model to the data by penalizing the log-likelihood value with respect to the estimated parameters. In all measures, a lower value means a better fit. We varied the number of classes from 2 to 5 and recorded the associated fit measures. We also empirically assessed the extent to which the resulting classes could be logically described and interpreted. Based on this, the four-class model did fit the best with the following membership proportions: $0.31, 0.28, 0.22,$ and 0.19.

4. Results and discussions

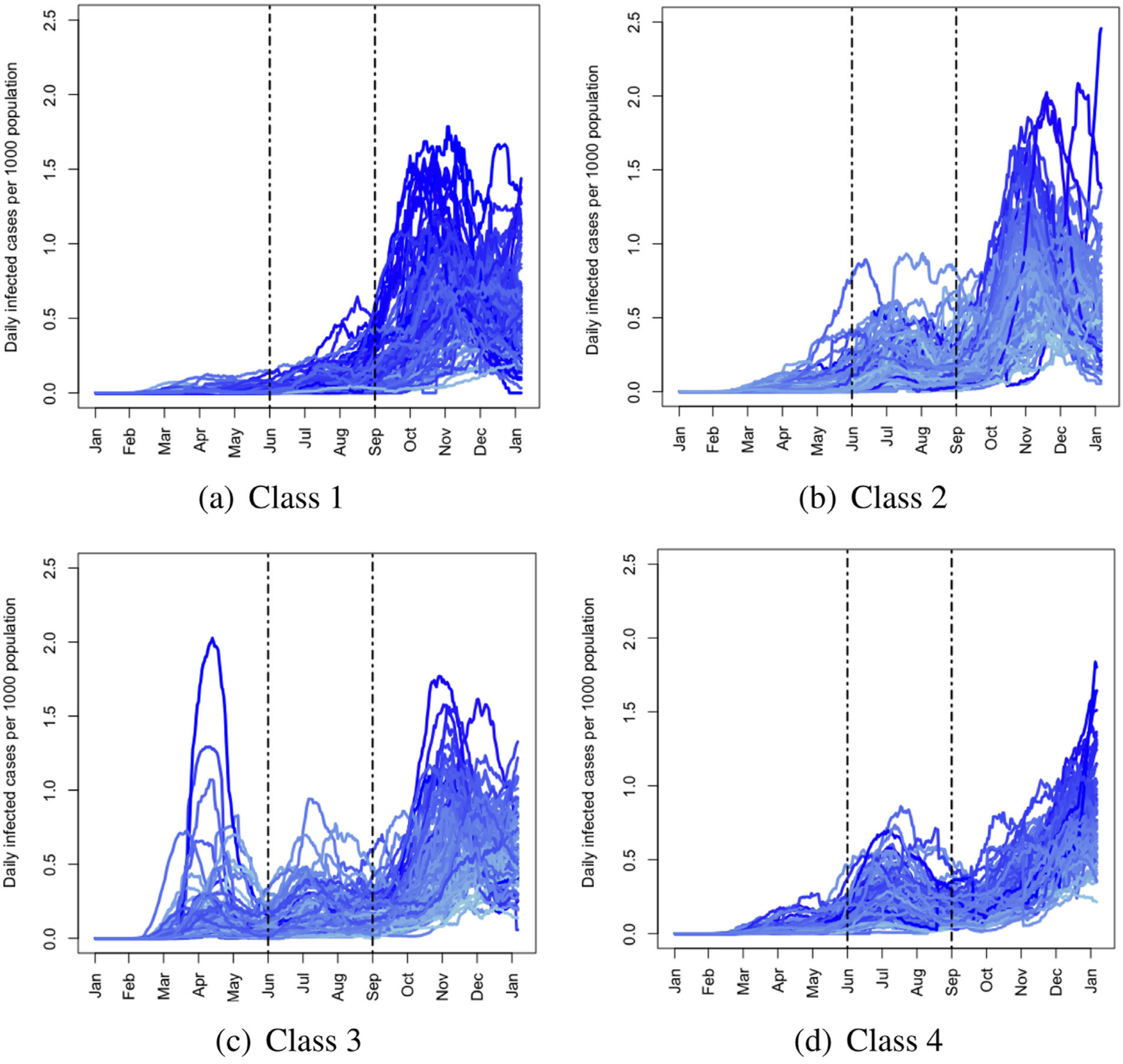
4.1. Four classes and their wave patterns

The class-conditional membership probabilities are shown in Table 3. The daily median infection rates for different classes are shown in Fig. 4a, where a county is assigned to the class that has the largest posterior probability (modal assignment). The corresponding spread patterns (the moving average of daily infections) for a randomly selected 100 counties from each given class are shown in Fig. 5.

The first class (Class 1, the largest class with 31 percent share) corresponds to the late-wave pattern that, as the name suggests,

Table 3
Class conditional probabilities of indicator variables (π-table).

	Class 1	Class 2	Class 3	Class 4
Class size (out of 3142)	**972**	**891**	**701**	**578**
Class probabilities	**0.31**	**0.28**	**0.22**	**0.19**
Early wave?				
Yes	0.00	0.00	**1.00**	0.00
No	**1.00**	**1.00**	0.00	1.00
Mid wave?				
Yes	0.00	**1.00**	**0.47**	1.00
No	**1.00**	0.00	0.53	0.00
Late wave or rise?				
Had wave (rise & fall)	**0.77**	**0.85**	**0.55**	0.00
Had only rise	0.20	0.13	0.45	**1.00**
No wave or rise	0.03	0.00	0.00	0.00
Height of early wave				
Wave does not exist	**1.00**	**1.00**	0.00	1.00
Low	0.00	0.00.	0.33	0.00
Medium	0.00	0.00	0.33	0.00
High	0.00	0.00	0.34	0.00
Height of mid wave				
Wave does not exist	**1.00**	0.00	**0.53**	0.00
Low	0.00	**0.39**	0.16	0.24
Medium	0.00	0.30	0.16	0.39
High	0.00	0.31	0.16	**0.37**
Height of late wave				
Neither wave nor rise	0.03	0.00	0.00	0.00
Low	**0.37**	0.27	**0.37**	**0.45**
Medium	0.28	0.35	0.33	0.33
High	0.33	**0.39**	0.30	0.22
Number of waves				
Zero	0.22	0.00	0.00	0.00
One	**0.77**	0.00	0.17	**0.88**
Two	0.01	**0.92**	**0.54**	0.11
Three or more	0.00	0.08	**0.28**	0.01
Metropolitan status				
Metro area	0.31	0.27	**0.44**	**0.55**
Non-metro area	**0.69**	**0.73**	0.56	0.45



(a) Class 1

(b) Class 2

(c) Class 3

(d) Class 4

Fig. 5. Pandemic wave patterns of four classes. (a) Class 1 has one late wave, (b) Class 2 has mid and late waves, (c) Class 3 have mixture of waves in all windows, (d) Class 4 has mid wave and late rise.

comprises a single wave only during the third pandemic window (77 percent, cf. Table 3). The counties belonging to this class experienced the pandemic hits comparatively later in the year and they do not contain any wave in the first and second windows (Fig. 5a). These counties also have varying heights of their late waves; a higher fraction of them (37 percent) had lower daily infection rates (low-height category). In terms of demographic characteristics, Class 1 counties mostly belong to non-metropolitan areas (69 percent). Prior studies also found that COVID spread was slow, at least at the beginning of the pandemic, in rural counties compared to the urban ones in the US [15,40].

The second class (Class 2) represents the mid and late-wave pattern containing counties that experienced two and more pandemic waves started from summer (mid-wave) to fall period (late-wave) (92 percent counties have two waves and the rests have three or more waves). We observe from Fig. 5b that, a higher fraction of counties of this class (39 percent) had low infection rates during the summer, but the rate subsequently got increased afterward and a higher fraction of counties (39 percent) fall into the higher range of infection rates in their late waves. Similar to Class 1, the majority of the counties (73 percent) of this class are from non-metropolitan areas.

The third class (Class 3) of counties (with a share of 22 percent out of 3142 counties) have a unique aspect that these counties have pandemic waves in all windows (early, mid, and late) in different proportions (cf. Table 3). This is also the only class that experienced a pandemic wave in the early phase of the pandemic (in the first window). Around 83 percent counties of this class had two or more number of waves throughout the year and 17 percent had only one wave (the one in the first window). The trend patterns shown in Fig. 5c reveal that Class 3 counties have higher infection rates in the early stage. The infection cases got flatten down a little bit in the middle of the year and then rose again the latter stage. The higher median values of new cases in the early window and the lower median values of cases in the mid window reveal this trend in Fig. 4b. Unlike Class 1 and Class 2, a considerable fraction of counties (above 44 percent) in this class are located in metropolitan areas.

Finally, the last class (Class 4) is the smallest class among the four classes (19 percent counties). The counties belonging to this

Table 4
Coefficients of active covariates with respect to class 1 (η-table).

Active covariates	Class 2	Class 3	Class 4
Intercept	3.016	2.316	-0.681
Frac. of people aged 60+	-4.390***	-5.268***	-3.239**
Frac. of African-Americans	2.385***	6.789***	6.159***
Social distance index			
In early wave	-0.039**	0.037**	0.022
In mid wave	0.003	0.045*	0.164***
In late wave	0.009	-0.061**	-0.172***
Test done per 1K (in log)			
In early wave	1.923***	1.087	-0.244
In mid wave	-1.571	-1.990	-7.231***
In late wave	-0.064	0.637	5.690***
Imported cases (in log)			
In early wave	-0.733***	1.133***	-0.437*
In mid wave	2.933***	0.828**	7.339***
In late wave	-2.178***	-1.734***	-5.793***

*, **, *** mean 10%, 5% and 1% level of significance respectively.

Table 5
Class-wise probability-weighted summary for covariates.

Active covariates	Class 1	Class 2	Class 3	Class 4
People aged 60+ (%)	27.0	25.0	24.0	24.0
African-Americans (%)	3.0	6.0	13.0	19.0
Social distance index	27.3	26.8	28.3	28.1
Imported cases (early)	3151	2649	17649	4881
Inactive covariates				
Population density	85.7	101.6	547.7	260.5
Cumulative inf. cases	81.9	108.3	115.9	101.3
Work % change	-22.5	-22.6	-24.9	-24.3
Non-work % change (early)	-11.8	-7.8	-16.3	-11.3
Non-work % change (mid)	12.3	4.3	-1.6	-7.4
Non-work % change (late)	-12.9	-15.6	-18.9	-16.8

class had no early waves, but had mid waves and late rises. These counties experienced higher infection rate (higher peak) during the second pandemic window (mid-wave) compared to other classes (cf. Fig. 4b). In the third window, however, the trend got only a rise but no fall (the infection cases are monotonically growing without any decline). The presence of this "rise" pattern in the late window is a unique aspect of this class compared to the other three classes. The monotone rise of median cases of this class (in Fig. 4b) is suggestive to this trend. Notably, a higher fraction of the counties of this class (55 percent) belong to metropolitan areas.

4.2. Prediction of latent class membership

The effects of active covariates on class membership are presented in Table 4. Here the coefficients (η_c's) relative to Class 1 are displayed. The class-wise probability weighted mean values for both active and inactive covariates are shown in Table 5. This table shows that Class 1 and Class 2 counties have lower population densities compared to Class 3 and 4. Since Class 1 counties experienced only the late pandemic waves and they have considerably low infection rates in early and mid year, these counties had the smallest cumulative infection cases across all classes, whereas Class 3 counties had the largest cumulative cases because they observed pandemic waves in all three windows.

Table 4 reveals that counties with higher fraction of people aged over 60 years are more likely to belong to Class 1 compared to other three classes. It indicates that the presence of higher fraction of older adults does not make a county more vulnerable for pandemic spread. Findings from other studies support this observation with the claim that people in the older age group have a lower tendency to spread the disease than the younger population [41,42]. In contrast, counties with a higher share of African-American people are less likely to belong to Class 1 than other classes. That means counties with a higher cumulative infection rate have a higher chance to have African-American people in their counties than the counties with low rates. The positive association between the infection rate and this racial group is also observed in [15,42–45].

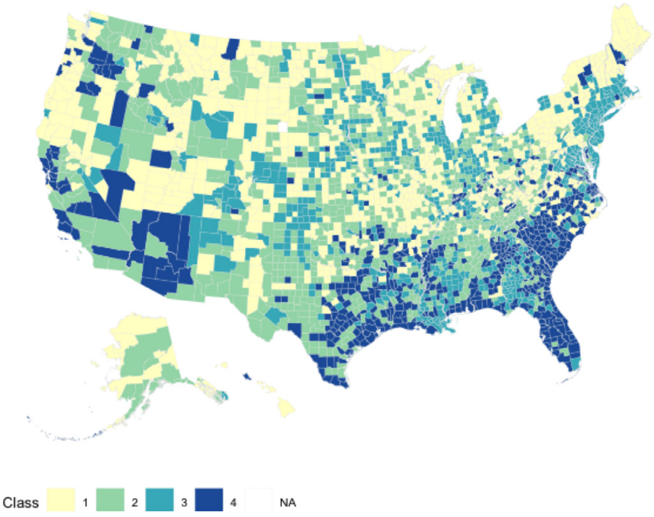


Fig. 6. Distribution of four latent classes across the US.

We observe from Table 5 that, during the early phase of the pandemic, counties belong to Class 3 and Class 4 got a considerably higher number of daily external trips from out of state/county (higher fraction of counties with high imported COVID cases). In addition, these two classes experienced higher reductions in work and non-work place visits than the baseline (Jan 3 - Feb 6, 2020) compared to Class 1 and Class 2. Notably, during summer (mid-wave), Class 1 and Class 2 counties experienced increased visits in various non-work places than the baseline (+ve percentage changes). On the other hand, Class 3 and Class 4 counties that are comparatively higher density counties posed restrictions on peoples mobility due to higher COVID spread, thus experienced lower non-work place visits than the baseline during that period (-ve percentage changes).

Higher social distance index in early phases of the pandemic posed a higher possibility of counties to be included in Class 1 than Class 2 (cf. Table 4). Although pandemic started late in Class 1 counties but those counties might follow the strict preventive measures in early phases than Class 2. On the other hand, a higher social distance index in the early phases increased the chance of counties to be included in Class 3 than Class 1. This might be due to the fact that the pandemic hit earlier in Class 3 counties than Class 1 that bound people to stay at home by imposing restrictions on the movement of people in order to contain the spread. Again, the higher social distance index during summer (the second window) increased the chances of counties to be included in Class 3 and Class 4 than Class 1. Furthermore, the higher tests done per population in the early window increases the propensity to belong to Class 2 and the higher testing during the later phase increases the chance of counties to belong to Class 4 compared to Class 1.

Figure 6 shows US county-level map by four identified classes and Table 6 lists the distribution of counties per class across four major US regions (the leading state per region is in parentheses). We observe that a higher number of Northeast counties experienced early pandemic hit (Class 3), whereas only a few West counties experienced the same. Most of the West counties had pandemic waves in the mid and late windows (Class 1 and 2). Most of the Midwest counties had a full wave only in the third window

Table 6
Distribution of latent classes across four US regions.

Region (#counties)	Class 1 (972)	Class 2 (891)	Class 3 (701)	Class 4 (578)
Northeast	109 (PA)	8 (PA)	94 (NY)	6 (NY)
Midwest	445 (MO)	334 (KS)	255 (IN)	21 (IN)
South	227 (KY)	407 (TX)	292 (GA)	496 (TX)
West	191 (MT)	142 (CA)	60 (CO)	55 (CA)

(Class 1) whereas South counties experienced a rising trend in the same window (Class 4).

5. Conclusions

This paper analyzes the heterogeneity of COVID-19 spread patterns across US counties and their association with distinct socio-demographic, location, and human mobility characteristics of those counties during the one year of the pandemic from January 2020 to January 2021. In particular, we at first characterize the pandemic spread patterns of counties and then cluster the counties into a set of distinct classes based on the similarity of their spread patterns. Results suggest that 3142 US counties can be divided into four such classes where the counties in each identified class have particular age and racial distribution, metropolitan status, social distance index, external trips, and other characteristics. For example, Class 1 is composed of non-metropolitan counties with a higher fraction of older adults (age 60+) and a lower fraction of African-Americans. These counties experienced only the fall pandemic waves and had the lowest number of external trips and cumulative infection cases across all classes. The majority of Class 2 counties also belong to non-metropolitan areas but unlike Class 1, this class experienced two waves during summer and fall periods. In contrast, a higher fraction of Class 3 and Class 4 counties represent metropolitan areas with high population density. These two classes had a higher number of external trips from outside counties/states and had higher reductions in both work and non-workplace visits during the pandemic. Unlike other classes, Class 3 counties had pandemic waves in all three periods: spring, summer, and fall. Last, Class 4 counties experienced summer wave and fall rise.

This study provides important insights regarding the variations in COVID-19 spread patterns across US counties during the first year of the pandemic. We note that immediately after our study period (that is, after January 2021), vaccinations efforts have been extensively rolled out throughout the US that helped COVID-19 cases decline drastically in many counties. We next plan to study the effect of vaccinations and their aftermath on spread patterns across US counties. Furthermore, in the future, we plan to develop a county-level prediction model for each identified class that would forecast the progress and evolution of COVID-19 spread for that class. Unlike a single model for all areas or counties, we argue that the class-specific modeling will ensure a better fitting of data and consequently will deliver better precision for future prediction, both short-term forecast, and long-term prediction. In that, local policymakers can use the developed models to track the spread of COVID-19 in their respective communities and can take necessary action items, such as preventive measures and resource allocations, to curve and contain the spread.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] COVID-19 coronavirus pandemic, 2022, (https://www.worldometers.info/coronavirus, accessed on June 9, 2022).
[2] B. Mishra, B. Mishra, G. Mishra, N.D. Sinha, Pandemic dynamics, the three waves of COVID-19 and the way ahead, J. Adv. Res. Med. Sci. Technol. 8 (1) (2021) 13–18. (ISSN: 2394–6539)
[3] A.G. Hoen, T.J. Hladish, et al., Epidemic wave dynamics attributable to urban community structure: a theoretical characterization of disease transmission in a large network, J. Med. Internet Res. 17 (7) (2015) e169.
[4] Y.-H. Hsieh, C. Chen, Turning points, reproduction number, and impact of climatological events for multi-wave dengue outbreaks, Trop. Med. Int. Health 14 (6) (2009) 628–638.
[5] L. Simonson, Pandemic influenza and mortality: past evidence and projections for the future, 2004.
[6] N.P. Johnson, J. Mueller, Updating the accounts: global mortality of the 1918–1920 “Spanish” influenza pandemic, Bull. Hist. Med. (2002) 105–115.
[7] V. Andreasen, C. Viboud, L. Simonsen, Epidemiologic characterization of the 1918 influenza pandemic summer wave in copenhagen: implications for pandemic control strategies, J. Infect. Dis. 197 (2) (2008) 270–278.
[8] E. Kaxiras, G. Neofotistos, Multiple epidemic wave model of the COVID-19 pandemic: modeling study, J. Med. Internet Res. 22 (7) (2020) e20912.
[9] COVIDTeam IHME, R. Reiner, R. Barber, J. Collins, Modeling COVID-19 scenarios for the United States, Nat. Med. (2020).
[10] S. Chang, E. Pierson, P.W. Koh, J. Gerardin, B. Redbird, D. Grusky, J. Leskovec, Mobility network models of COVID-19 explain inequities and inform reopening, Nature 589 (7840) (2021) 82–87.
[11] I. Nesteruk, Waves of COVID-19 pandemic. detection and sir simulations, medRxiv (2020).
[12] A.J. Kucharski, T.W. Russell, et al., Early dynamics of transmission and control of COVID-19: a mathematical modelling study, Lancet Infect. Dis. 20 (5) (2020) 553–558.
[13] J.S. Jia, X. Lu, Y. Yuan, G. Xu, J. Jia, N.A. Christakis, Population flow drives spatio-temporal distribution of COVID-19 in China, Nature 582 (7812) (2020) 389–394.
[14] R. Rui, M. Tian, M.-L. Tang, G.T.-S. Ho, C.-H. Wu, Analysis of the spread of COVID-19 in the usa with a spatio-temporal multivariate time series model, Int. J. Environ. Res. Public Health 18 (2) (2021) 774.
[15] R. Paul, A.A. Arif, O. Adeyemi, S. Ghosh, D. Han, Progression of COVID-19 from urban to rural areas in the United States: a spatiotemporal analysis of prevalence rates, J. Rural Health 36 (4) (2020) 591–601.
[16] G. Soldi, N. Forti, D. Gaglione, P. Braca, L.M. Millefiori, S. Marano, P.K. Willett, K.R. Pattipati, Quickest detection and forecast of pandemic outbreaks: analysis of COVID-19 waves, IEEE Commun. Mag. 59 (9) (2021) 16–22.
[17] I. Nesteruk, Comparison of the first waves of the COVID-19 pandemic in different countries and regions, in: COVID-19 Pandemic Dynamics, Springer, 2021, pp. 89–107.
[18] M. Rusiñol, I. Zammit, M. Itarte, E. Forés, S. Martínez-Puchol, R. Girones, C. Borrego, L. Corominas, S. Bofill-Mas, Monitoring waves of the COVID-19 pandemic: inferences from WWTPs of different sizes, Sci. Total Environ. 787 (2021) 147463.
[19] L. Rampal, B. Liew, M. Choolani, K. Ganasegeran, A. Pramanick, S. Vallibhakara, P. Tejativaddhana, V. Hoe, Battling COVID-19 pandemic waves in six south-east Asian countries: a real-time consensus review, Med. J. Malaysia 75 (6) (2020) 613–625.
[20] G. Cacciapaglia, C. Cot, F. Sannino, Second wave COVID-19 pandemics in europe: a temporal playbook, Sci. Rep. 10 (1) (2020) 1–8.
[21] P. Domingo, V. Pomar, I. Mur, I. Castellví, H. Corominas, N. de Benito, Not all COVID-19 pandemic waves are alike, Clin. Microbiol. Infect. 27 (7) (2021) 1040–e7.
[22] R.C. O'Connor, K. Wetherall, S. Cleare, H. McClelland, A.J. Melson, C.L. Niedzwiedz, R.E. O'Carroll, D.B. O'Connor, S. Platt, E. Scowcroft, et al., Mental health and well-being during the COVID-19 pandemic: longitudinal analyses of adults in the UK COVID-19 mental health & wellbeing study, Br. J. Psychiatry 218 (6) (2021) 326–333.
[23] W. Rothengatter, J. Zhang, Y. Hayashi, A. Nosach, K. Wang, T.H. Oum, Pandemic waves and the time after COVID-19–consequences for the transport sector, Transp. Policy 110 (2021) 225–237.
[24] T. Hale, N. Angrist, A.J. Hale, B. Kira, S. Majumdar, A. Petherick, T. Phillips, D. Sridhar, R.N. Thompson, S. Webster, et al., Government responses and COVID-19 deaths: global evidence across multiple pandemic waves, PLoS ONE 16 (7) (2021) e0253116.
[25] R.A. Rios, T. Nogueira, D.B. Coimbra, T.J. Lopes, A. Abraham, R.F.d. Mello, Country transition index based on hierarchical clustering to predict next COVID-19 waves, Sci. Rep. 11 (1) (2021) 1–13.
[26] S.X. Zhang, F.A. Marioli, R. Gao, S. Wang, A second wave? what do people mean by COVID waves?–A working definition of epidemic waves, Risk Manage. Healthc Policy 14 (2021) 3775.
[27] T. Fisayo, S. Tsukagoshi, Three waves of the COVID-19 pandemic, Postgrad. Med. J. 97 (1147) (2021) 332.
[28] T. Zhang, G. Lin, Generalized k-means in GLMs with applications to the outbreak of COVID-19 in the United States, Comput. Stat. Data Anal. 159 (2021) 107217.
[29] J.A. Hagenaars, A.L. McCutcheon, Applied Latent Class Analysis, Cambridge University Press, 2002.
[30] John Hopkins University, COVID-19 data repository dataset, 2020.

[31] Maryland Transportation Institute University of Maryland, University of Maryland COVID-19 Impact Analysis Platform, 2020. https://data.covid.umd.edu.

[32] Google, Google COVID-19 community mobility reports, 2020.

[33] U.C.B. 2018, 2014-2018 american community survey 5-year estimates [www document], 2018, http://www.census.gov/programs-surveys/acs/news/data-releases/2018/release.html (accessed 11.23.20).

[34] D.D. Ingram, S.J. Franco, NCHS urban-rural classification scheme for counties, Vital Health Stat Ser. 2 2 (154) (2012) 1–65.

[35] J.K. Vermunt, J. Magidson, Latent class cluster analysis, Appl. Latent Class Anal. 11 (89–106) (2002) 60.

[36] S.T. Lanza, B.L. Rhoades, Latent class analysis: an alternative perspective on subgroup analysis in prevention and treatment, Prev. Sci. 14 (2) (2013) 157–168.

[37] D.A. Linzer, J.B. Lewis, et al., poLCA: an R package for polytomous variable latent class analysis, J. Stat. Softw. 42 (10) (2011) 1–29.

[38] H. Akaike, Information Theory and an Extension of the Maximum Likelihood Principle, in: Selected papers of Hirotugu Akaike, Springer, 1998, pp. 199–213.

[39] D. Oberski, Mixture models: latent profile and latent class analysis, in: Modern Statistical Methods for HCI, Springer, 2016, pp. 275–287.

[40] New-York-Times, Coronavirus was slow to spread to rural America, not anymore, 2020,

[41] A.M. Oster, Transmission dynamics by age group in COVID-19 hotspot counties united states, april–september 2020, MMWR Morb. Mortal. Wkly. Rep. 69 (2020) 1494–1496.

[42] R. Rafiq, T. Ahmed, M.Y.S. Uddin, Structural modeling of COVID-19 spread in relation to human mobility, Transp. Res. Interdiscip. Perspect. 13 (2022) 100528.

[43] R. Rafiq, M.G. McNally, Y.S. Uddin, T. Ahmed, Impact of working from home on activity-travel behavior during the COVID-19 pandemic: an aggregate structural analysis, Transp. Res. Part A PolicyPract. 159 (2022) 35–54.

[44] R. Kullar, J.R. Marcelin, T.H. Swartz, D.A. Piggott, R. Macias Gil, T.A. Mathew, T. Tan, Racial disparity of coronavirus disease 2019 in African American communities, J. Infect. Dis. 222 (6) (2020) 890–893.

[45] D.B.G. Tai, A. Shah, C.A. Doubeni, I.G. Sia, M.L. Wieland, The disproportionate impact of COVID-19 on racial and ethnic minorities in the United States, Clin. Infect. Dis. (2020).