Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Epidemics 41 (2022) 100640

Contents lists available at ScienceDirect

Epidemics

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

A comparison of COVID-19 outbreaks across US Combined Statistical Areas using new methods for estimating R_0 and social distancing behaviour

Ludovica Luisa Vissat [a], Nir Horvitz [a], Rachael V. Phillips [b], Zhongqi Miao [a], Whitney Mgbara [a], Yue You [b], Richard Salter [c], Alan E. Hubbard [b], Wayne M. Getz [a,b,d,*]

[a] Department of Environmental Science, Policy, and Management, UC Berkeley, CA 94720, USA
[b] Division Environmental Health Sciences, UC Berkeley, CA 94720, USA
[c] Computer Science Department, Oberlin College, Oberlin, Ohio, OH 44074, USA
[d] School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban 4000, South Africa

ARTICLE INFO

Keywords:
COVID-19
Curve flattening index
Epidemiological modelling
R_0
Social distancing

ABSTRACT

We investigated the initial outbreak rates and subsequent social distancing behaviour over the initial phase of the COVID-19 pandemic across 29 Combined Statistical Areas (CSAs) of the United States. We used the Numerus Model Builder Data and Simulation Analysis (NMB-DASA) web application to fit the exponential phase of a SCLAIV+D (Susceptible, Contact, Latent, Asymptomatic infectious, symptomatic Infectious, Vaccinated, Dead) disease classes model to outbreaks, thereby allowing us to obtain an estimate of the basic reproductive number R_0 for each CSA. Values of R_0 ranged from 1.9 to 9.4, with a mean and standard deviation of 4.5 ± 1.8. Fixing the parameters from the exponential fit, we again used NMB-DASA to estimate a set of social distancing behaviour parameters to compute an epidemic flattening index c_{flatten}. Finally, we applied hierarchical clustering methods using this index to divide CSA outbreaks into two clusters: those presenting a social distancing response that was either weaker or stronger. We found c_{flatten} to be more influential in the clustering process than R_0. Thus, our results suggest that the behavioural response after a short initial exponential growth phase is likely to be more determinative of the rise of an epidemic than R_0 itself.

1. Introduction

The two epidemiological statistics that best characterise the seriousness of any epidemic are the initial reproductive number R_0 (R-zero; the expected number of new cases arising from the initiating or patient zero case) and the disease-related mortality rate α (mathematical epidemiologists call this the virulence parameter, Anderson and May 1992). These statistics, however, depend both on intrinsic biological factors (pathogen invasion and host response biology) and socio-environmental factors (contact behaviour of hosts and pathogen survival in the environment) (Delamater et al., 2019), but they only convey part of the understanding of potential of an epidemic to cause harm. The other part is how easily R_0 and α can be reduced through changes in host behaviour (Davies et al., 2020) and, in the case of humans, the use of pharmaceuticals and treatments to reducing both the disease mortality rate α (Ali et al., 2020a; Tsang et al., 2021) and the time-varying reproductive value $R_{\text{eff}}(t)$ (R-effective, where $R_{\text{eff}}(0) \equiv R_0$; Scherer and McLean 2002) as the epidemic unfolds.

For completeness we briefly note that in epidemics where infected individuals become immune after recovering from their illness, $R_{\text{eff}}(t)$ reduces from its initial value R_0, which must exceed 1 for the epidemic to occur, until it reaches the value $R_{\text{eff}}(t) < 1$. At this point, the epidemic begins to shut down because fewer new cases occur in the next generation of infected individuals than in the current generation. This takes place when the number of susceptible individuals falls below a critical proportion of individuals in the population because many of the individuals coming into contact with infectious individuals have previously had the disease and are now immune or social distancing behaviour has reduced contact rates between susceptible (S) and infectious (I) individuals.

Of the two statistics, the reproductive number R_0 is much more difficult to measure than the disease-induced mortality rate α. The latter requires that we only keep track of those dying from the disease over time and then compute an average number dying per unit time.

Abbreviations: CSA, Combined Statistical Area; NMB-DASA, Numerus Model Builder Data and Simulation Analysis; SCLAIV+D, susceptible-contact-latent-asymptomatic infectious-symptomatic infectious-vaccinated-dead

* Corresponding author at: Department of Environmental Science, Policy, and Management, UC Berkeley, CA 94720, USA.
E-mail address: wgetz@berkeley.edu (W.M. Getz).

https://doi.org/10.1016/j.epidem.2022.100640
Received 15 November 2021; Received in revised form 3 October 2022; Accepted 3 October 2022
Available online 10 October 2022

Computation of R_0 requires that we estimate both the average number of individuals that each infected individual infects per unit time at the start of an outbreak and the average time between generations of infected individuals (the first generation of infected individuals are those infected by patient zero and the second generation are those infected by the first generation, and so on for subsequent generations; also see Lehtinen et al. 2021). Further, the most accurate way to assess who infected who and the time between when the donor and recipient individuals were first infected involves some level of contact tracing, which formerly labour intensive is now becoming easier to accomplish with the advent of digital and social media technologies (Anglemyer et al., 2020).

The importance of the statistic R_0 to epidemiological community can be seen by the fact that considerable effort went into reporting its value early into the COVID-19 pandemic. Yet, as our results will suggest, evaluating the behavioural response of individuals to the epidemic may be more determinative of the rise of the epidemic than R_0. Clearly, however, other factors start to become important as well, including non-pharmaceutical intervention (NPI) stringency and lockdown measure (Kishore et al., 2022; Kraemer et al., 2020). A systematic review of a literature search conducted on September 15, 2020 (Billah et al., 2020), identified 42 papers published in English that reported estimates of R_0 for COVID-19 ranging from a high of 6.91 to values less than 1. Early into the pandemic initial estimates of R_0 (Li et al., 2020; Wu et al., 2020) suggested that its value was in the range 2–3, a similar range to that estimated for the original SARS outbreak (SARS-CoV-1 pathogen) in 2002 to 2003 (Anderson et al., 2004). Since SARS-CoV-1 was contained, by June, 2003 – with a reported 8098 cases, most acquired nosocomially, and 774 deaths (Petersen et al., 2020) – SARS-CoV-2 did not initially raise major concerns regarding a potential global spread. Other studies soon thereafter showed that R_0 was actually much higher (Sanche et al., 2020; Liu et al., 2020) and therefore epidemiologists realised that COVID-19 (SARS-CoV-2 pathogen) was going to be much more challenging to manage than the original SARS outbreak. Equally, epidemiologists became alarmed with the emergence of the Delta variant of SARS-CoV-2, when it was shown that this strain was estimated to have an R_0 of nearly twice that of the early SARS-CoV-2 strains: 2.79 versus 5.08 (Liu and Rocklöv, 2021).

The fact that values of R_0 are often cited to two decimal places belies the reality that the same pathogen invading two different communities can result in estimates of R_0 that differ by several fold. Such a range of R_0 estimates for COVID-19 have been presented in meta-analyses and literature reviews (Billah et al., 2020; Kochańczyk et al., 2020), as well as in the study reported here. Wide variation may be due to two factors: (i) estimating R_0 is not a statistically robust process, given the small numbers of infected individuals that occur at the start of an outbreak as well as the limited availability of outbreak data; and (ii) various socio-environmental processes influence the course of an outbreak, including whether or not it is likely to occur in the first place.

A number of statistical studies have been undertaken to elucidate the socio-environmental factors that influence the severity of COVID-19 in the US and around the world (Mehmood et al., 2021; Rozenfeld et al., 2020; Wang et al., 2020; Zhang et al., 2021). These include identifying individual host traits – such as race, ethnicity, income level, age group, use of public transport, and so on – that place individuals at higher risk for infection with COVID-19 (Chen and Krieger, 2021; McCoy et al., 2021; Torrats-Espinosa, 2021; Voinsky et al., 2020). They also include relating measures obtained from empirical data on the severity of a COVID-19 outbreak in a local community with community measures of its age-structure, ethnic and racial make up, use of public transport (McCoy et al., 2021), climatic factors (Briz-Redón and Serrano-Aroca, 2020), and even air pollution (Travaglio et al., 2021). For example, Siedner et al. (2020) looked at mortality growth rate to understand the effects of social distancing response, Khataee et al. (2021) estimated mitigation efforts from mobile phone tracking data, Weill et al. (2020) showed how individuals in wealthier areas had decreased their mobility significantly more than individuals in poorer areas, and Grantz et al. (2020) showed how mobile phones can be used to collect population response data. These studies, however, relied on empirical social distancing or mobility response data to obtain their results. Such data may not be readily available or even exist.

Here we take a different and novel tack. We obtain social behavioural measures from epidemiological models fitted to empirical data, rather than from the empirical data itself. Deriving these measures from a model requires incidence data only. The actual social distancing behaviour is inferred from fitting an epidemiological model that includes social distancing effects to incidence data (Getz et al., 2021). The data we use are reported at the level of Combined Statistical Areas (CSAs) by the United States Office of Management and Budget (Management and Budget Office, 2010).

CSAs, of which there are 172 in the US, provide comprehensive, coherent population units that have a relatively high degree of socioeconomic integration of the metropolitan and micropolitan areas contained within their boundaries (see Section 2 for definitions). More, specifically, metropolitan areas within a particular CSA are typically linked by the same commuter rail service and served by the same major commercial airport (Ottensmann, 2017). Hence CSAs provide a better fit than metropolitan or micropolitan areas due to the modelling assumption of an epidemic occurring in a homogeneous, well-mixed population that is relatively isolated from its surroundings (i.e., many individuals in the same CSA move daily among their constituent metropolitan and micropolitan areas, while the same cannot be said of neighbouring CSA units). Lack of complete isolation is then dealt with by analysing collections of populations in a metapopulation setting (Keeling et al., 2004).

We were particularly interested in extracting information related to the initial exponential outbreak phase and the social distancing response within different CSAs, and consequently to compare these responses across CSAs. The NMB-DASA web app (Getz et al., 2021) provides us with a tool for performing this kind of analysis. The underlying epidemiological model (SCLAIV+D model, described in more detail in Section 3) explicitly divides the population into basic (non-response) and response classes, allowing us to evaluate the response behaviour at a population level, as measure over a designated period of time.

Although we apply our framework to analysing the initial outbreak and response phase of the COVID-19 epidemic across CSAs in the US, it can just as easily be applied to other directly transmissible disease, such as influenza, measles, and the like. For the edification of our reader, we provide the following road map to our study, with details reported in the remaining sections of this paper. In summary, we:

1. identified the most suitable of the 172 CSAs in the US in terms of cases and mortality rates (data downloaded from https://usafacts.org/);
2. determined an approximation of the start date of the exponential outbreak phase for each CSA;
3. used the SCLAIV+D model of the NMB-DASA web app to extract the epidemic's initial exponential growth rate r in the 29 CSAs that made our suitability cut;
4. used Rai et al. (2020)'s best estimate of the generation time $\mathcal{T}$ (also serial interval—see Lehtinen et al. 2021) and fitted values of r to estimate R_0 for each of the 29 CSAs;
5. used the SCLAIV+D model to extract a curve flatting index that pertains to the social-distancing response phase in each of the 29 CSAs;
6. applied hierarchical clustering to our estimates of curve flattening behaviour across the 29 CSAs to reveal a two cluster structure that we interpreted as low versus high social-distancing response categories.

Unfortunately, given the fact that only 29 of the 172 CSAs had outbreak data that made our choice of suitability cut for estimating the epidemics initial exponential growth rate r, our sample size was

insufficient to carry out multiple hypothesis tests regarding factors that could statistically differentiate between our low versus high social-distancing response categories (Dziak et al., 2012). It would have taken more than half of our CSAs to make our suitability cut before an even limited factor analysis could be made. Reasons for outbreaks in CSAs not making our cut could include a possible initial stuttering to the start of the outbreak (Blumberg and Lloyd-Smith, 2013), but more likely that the initial outbreak in a majority of communities was disruptingly fed by infected individuals entering the population rather than being generated through an autochthonous process that later took hold (Kraemer et al., 2020).

2. Combined statistical areas data

The US Office of Management and Budget organises demographic and social data by the following population groupings:

- *Micropolitan area:* a delineated settled concentration of 10,000 to 49,999 individuals.
- *Metropolitan area:* a delineated settled concentration of ≥ 50,000 individuals.
- *Core*: an area within a larger region that has been delineated as a micropolitan or metropolitan area.
- *Employment interchange measure*: the sum of the percentage of workers living in a smaller entity (e.g., micropolitan area, or small metropolitan area) who work in a larger entity (e.g., a large metropolitan area) and the percentage of employment in the smaller entity that is accounted for by workers who reside in the larger entity.
- *Core Based Statistical Area (CBSA)*: a geographic entity within which statistical data are aggregated; where this entity consists of one or more counties jointly containing at least one core area, plus adjacent counties that have a high degree of social and economic integration through commuting ties with the identified core.
- *Combined Statistical Area (CSA)*: a geographic entity consisting of two or more adjacent Core Based Statistical Areas with employment interchange measure of at least 15.

The CSA data we used pertains to the period January 21, 2020 (which is the date of first COVID-19 case in the US) to September 5, 2020. This period takes us through the first major wave of COVID-19 in the US, which occurred during the period mid June to early September. A second larger wave occurred from late September to mid March 2021, and factors driving these two waves may be somewhat different.

To identify the most important COVID-19 CSAs, we first ranked them by population size, then by number of cases per 1000 people, and finally by the probability of deaths per case. We then summed these 3 rankings and used the 32 CSAs with the lowest scores in our study because a sizeable gap occurred between the 32nd and 33rd positions scores (see SOF1 for details). With this procedure, we obtained the 32 CSAs listed in Fig. 1. After further trimming 3 of these CSAs for a reason described in subsequent sections, we finally obtained the 29 CSAs coloured in red in Fig. 1.

3. NMB-DASA web app and SCLAIV+D model

We used the NMB-DASA web app and the SCLAIV+D model to estimate the values of R_0, generate values for our measure of social distancing, and perform our response comparison across CSAs. Because levels of social distancing vary over time, we obtained a measure of the levels that occurred for several weeks immediately after it was recognised within the communities that an initial outbreak had occurred. The SCLAIV+D model, reported in Getz et al. (2021) and available for use as a web app at NMB-DASA, was specifically designed to generate an integrative social distancing measure over specified periods of time. The model can be directly fitted to one or both of incidence and mortality data. The model includes both basic (regular) and response (reduced transmission) groups in each of the S (susceptible), C (contact, with the possibility of transmission), L (equivalent to E in a SEIR model), A (infectious but asymptomatic), I (infectious and symptomatic), V (equivalent to R in a SEIR model but may include vaccinated individuals), and dead (D) diseases classes (Fig. 2). The individuals move between the basic SCLAIV and SCLAIV-response classes according to various response drivers. These drivers can be specified using a constant or a time-varying rate. In particular, they are (as shown in Fig. 2):

1. the level of surveillance (i.e., proportion of cases detected) implemented to monitor the state of the outbreak
2. a social distancing rate (transfers individuals from S to S_r)
3. a social relaxation rate (transfers individuals from S_r to S)
4. a quarantine rate linked to contact tracing (transfers individuals from C, L and A to C_r, L_r, and A_r respectively)
5. case isolation (transfers individuals from I to I_r)
6. contact reduction (transfers individuals from C to C_r)
7. vaccination rates (transfers individuals from pre-infectious classes to V)
8. the impacts of treatment on disease-induced mortality rates

More details regarding the model description can be found in Getz et al. (2021). Using incidence data as input, selected parameters can be fitted to the model using maximum-likelihood estimation (MLE) and social distancing, social relaxation, surveillance or virulence drivers can be extracted. In addition, the analysis of the model output allows us to study the temporal dynamics of the different classes of individuals.

4. Exponential phase of community transmission

From a dynamical system point of view, a disease outbreak in a focal population is a stochastic process in which one or more imported cases begins a transmission chain that either peters out or goes on to generate the exponential outbreak phase of an ensuing epidemic (Blumberg and Lloyd-Smith, 2013). Thus, extracting an outbreak phase starting date goes beyond simply choosing to start the fitting procedure from the date of the first recorded case, since this case may well be an imported case that does not generate an outbreak transmission chain.

To identify an approximation of the start date of the exponential outbreak phase for each CSA, we used the following novel procedure. We slid a 15-day exponential curve-fitting window over the first several weeks of incidence data to find the best 15-day fit to the empirical data appearing in each window. The best fit was given by the range that produced the smallest sum of the absolute values of the residuals, provided a threshold growth rate (equal to 0.2) was exceeded, as described in more detail in SOF1. The first day of this best 15-day fit was taken as the start of the *exponential outbreak phase* of the epidemic. Three CSAs (Corpus Christi-Kingsville-Alice, TX, McAllen-Edinburg, TX and Montgomery-Selma-Alexander City, AL) were excluded because they did not satisfy the chosen criterion for identifying a suitable start date. We depict the results of our starting data identification in Fig. 3.

5. SCLAIV+D model fitting procedures

We used the NMB-DASA Web App to fit the basic parameters, which are contact rate, initial contacted population (C), initial symptomatic infectious population (I) and the succumb period (how long it takes after contact to being exposed, since not all infectious contacts results in an exposure that leads on to a symptomatic or asymptomatic infectious period). The period of this fitting was the first 2 weeks (days 0–14) of data from the selected starting date (see Getz et al. 2021 and SOF1 for more details, while a guide to reproducing this analysis is presented in SOF2). We chose the fit that minimised the absolute log-likelihood error among the output of 200 runs. We then fixed the fitted basic parameters and captured the initial social distancing response by fitting the social

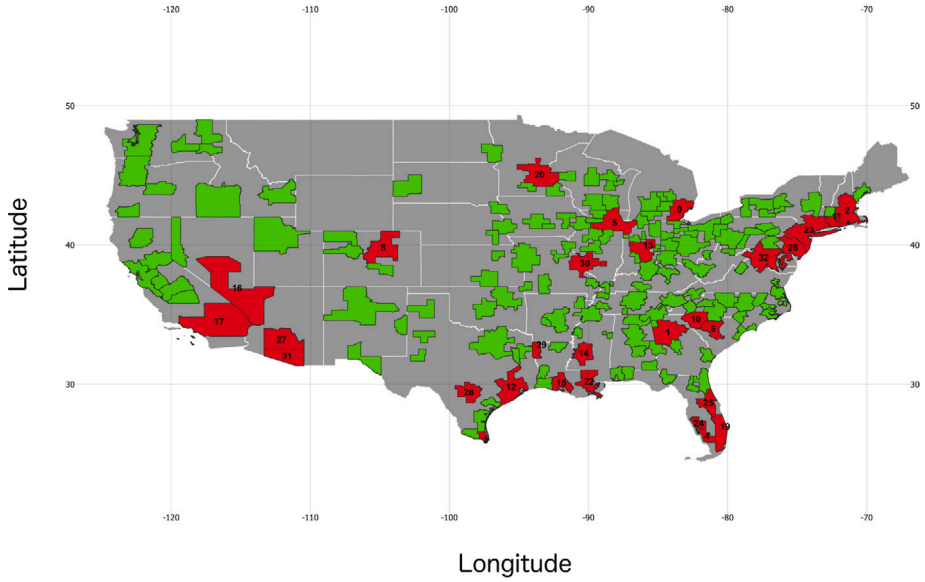

Fig. 1. The US Census bureau identifies 172 CSAs in the US (see text for the criteria used to identify CSAs). Of these, 32 listed here in this caption made the cut in our CSA ranking, while only 29 CSAs (red areas) met our criteria for inclusion in our analysis (those with sufficiently vigorous epidemics near the start of the COVID-19 pandemic) while 143 (green areas) did not. The 32 areas are: 1. Atlanta-Athens-Clarke County-Sandy Springs, GA-AL, 2. Boston-Worcester-Providence, MA-RI-NH-CT, 3. Brownsville-Harlingen-Raymondville, TX, 4. Cape Coral-Fort Myers-Naples, FL, 5. Chicago-Naperville, IL-IN-WI, 6. Columbia-Orangeburg-Newberry, SC, 7. Corpus Christi-Kingsville-Alice, TX, 8. Denver-Aurora, CO, 9. Detroit-Warren-Ann Arbor, MI, 10. Greenville-Spartanburg-Anderson, SC, 11. Hartford-East Hartford, CT, 12. Houston-The Woodlands, TX, 13. Indianapolis-Carmel-Muncie, IN, 14. Jackson-Vicksburg-Brookhaven, MS, 15. Lafayette-Opelousas-Morgan City, LA, 16. Las Vegas-Henderson, NV, 17. Los Angeles-Long Beach, CA, 18. McAllen-Edinburg, TX, 19. Miami-Port St. Lucie-Fort Lauderdale, FL, 20. Minneapolis-St. Paul, MN-WI, 21. Montgomery-Selma-Alexander City, AL, 22. New Orleans-Metairie-Hammond, LA-MS, 23. New York-Newark, NY-NJ-CT-PA, 24. North Port-Sarasota, FL, 25. Orlando-Lakeland-Deltona, FL, 26. Philadelphia-Reading-Camden, PA-NJ-DE-MD, 27. Phoenix-Mesa, AZ, 28. San Antonio-New Braunfels-Pearsall, TX, 29. Shreveport-Bossier City-Minden, LA, 30. St. Louis-St. Charles-Farmington, MO-IL, 31. Tucson-Nogales, AZ, 32. Washington-Baltimore-Arlington, DC-MD-VA-WV-PA.

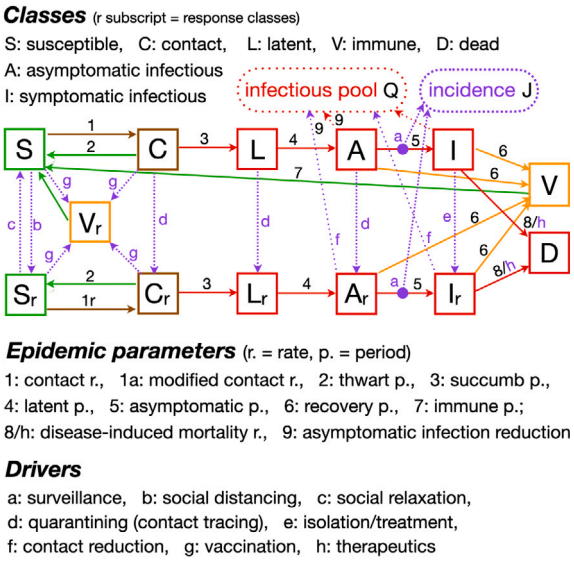

Fig. 2. A flow diagram of the SCLAIV+D model with its 8 flow parameters identified by numbers 1–8 plus an asymptomatic infection parameter 9. The 8 drivers identified by lower case Roman letters a-h are either zero (apart from surveillance), a positive constant or have the form of a switching function.
Source: Reprinted here from Getz et al. 2021.

distancing parameters over the longer period of day 0 to 30, repeating the procedure 200 times to find the best fit. The social distancing parameters that we extracted describe the time-varying driver as a switching function. In particular, they are: onset time, and the initial and final social distancing rates, as well as the switching time location, as explained in Getz et al. (2021). The results of these fittings are reported in SOF3, where the values of the best fitting parameters as well as the output of the model simulations using these parameters can be found.

6. R-zero estimation

Beyond estimating the start dates, the statistic that we were interested in extracting from our analysis was the estimated (best fitting) exponential growth rates r associated with the initial outbreaks of COVID-19 in each CSA. To obtain this value, we fitted an exponential curve to the trajectory produced by the SCLAIV+D model fitted to the incidence data over the first 2 weeks. We did this, rather than fitting an exponential curve directly to the incidence data itself, to obtain a growth rate estimation consistent with the set of basic parameters of the SCLAIV+D model that best fit the incidence curve to the first 2 weeks of empirical incidence data.

Our fitted exponential growth rate allows us to extract the transmission parameter β for given contact rate values, as described in Getz et al. (2021). In addition, a value for the basic reproduction number R_0 can be inferred from a value of r. This inference requires that we have an estimate $\mathcal{T}$ of the average time between infections, which is called the *generation time*. This time is on average equal to the serial interval time—which is the time from the onset of symptoms in an infector to the time of the onset of symptoms averaged across the one or more individuals (infectees) infected by this infector (Lehtinen et al., 2021). Note that, as shown in Ali et al. (2020b), generation time estimates are influenced by the isolation level of the transmission pairs. Strong isolation efforts would lead to a lower estimate of the generation time. Once both r and $\mathcal{T}$ are known, then R_0 can be computed from the relationship

$$R_0 = e^{r\mathcal{T}} \tag{1}$$

rather than the approximation $R_0 \approx 1 + r\mathcal{T}$ (Anderson et al., 2020) that can be used if $r\mathcal{T}$ is much smaller than 1. We note, however, that

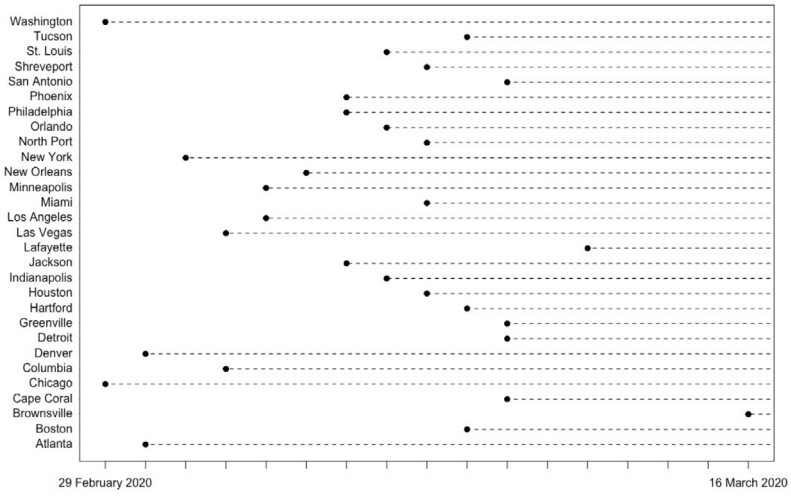

Fig. 3. Our computed day 0 for each CSA of interest. The earliest starting date we obtained was February 29, 2020, in the CSAs Washington-Baltimore-Arlington, DC-MD-VA-WV-PA and Chicago-Naperville, IL-IN-WI. The latest in our group was the Brownsville-Harlingen-Raymondville, TX CSA with a starting date of March 16, 2020. To keep our figure compact, we indicate the CSAs with the name of the first city in the full description provided in the caption of Fig. 1.

Table 1
Calculation of R_0 using Eq. (1) and generation time estimate $\mathcal{T} = 5.40$.

CSA name	r	R_0
Atlanta-Athens-Clarke County-Sandy Springs, GA-AL	0.26352	4.1495
Boston-Worcester-Providence, MA-RI-NH-CT	0.28437	4.6441
Brownsville-Harlingen-Raymondville, TX	0.20256	2.9856
Cape Coral-Fort Myers-Naples, FL	0.20758	3.0677
Chicago-Naperville, IL-IN-WI	0.38237	7.8837
Columbia-Orangeburg-Newberry, SC	0.24668	3.7889
Denver-Aurora, CO	0.25350	3.9310
Detroit-Warren-Ann Arbor, MI	0.23920	3.6389
Greenville-Spartanburg-Anderson, SC	0.19894	2.9279
Hartford-East Hartford, CT	0.25359	3.9328
Houston-The Woodlands, TX	0.33278	6.0316
Indianapolis-Carmel-Muncie, IN	0.34923	6.5918
Jackson-Vicksburg-Brookhaven, MS	0.23446	3.5470
Lafayette-Opelousas-Morgan City, LA	0.33598	6.1367
Las Vegas-Henderson, NV	0.25175	3.8941
Los Angeles-Long Beach, CA	0.24309	3.7160
Miami-Port St. Lucie-Fort Lauderdale, FL	0.26063	4.0854
Minneapolis-St. Paul, MN-WI	0.18176	2.6685
New Orleans-Metairie-Hammond, LA-MS	0.23877	3.6305
New York-Newark, NY-NJ-CT-PA	0.40881	9.0935
North Port-Sarasota, FL	0.21643	3.2179
Orlando-Lakeland-Deltona, FL	0.27489	4.4125
Philadelphia-Reading-Camden, PA-NJ-DE-MD	0.29393	4.8901
Phoenix-Mesa, AZ	0.41426	9.3651
San Antonio-New Braunfels-Pearsall, TX	0.21134	3.1306
Shreveport-Bossier City-Minden, LA	0.28708	4.7125
St. Louis-St. Charles-Farmington, MO-IL	0.29823	5.0051
Tucson-Nogales, AZ	0.12113	1.9234
Washington-Baltimore-Arlington, DC-MD-VA-WV-PA	0.26769	4.2442
MEAN ± SD	0.27 ± 0.07	4.5 ± 1.8
MEDIAN	0.25	3.93

Eq. (1) applies only when the generation time distribution is Gaussian-like, with notable errors occurring when the distribution is exponential (Wearing et al., 2005a).

Under the assumption that the generation time is the same for each CSA, the calculations of R_0 for each of the CSAs will vary with their estimated exponential growth rates r. In Table 1, we report our CSA values R_0, using a generation time $\mathcal{T} = 5.40$, as obtained by Rai et al. (2020) in their meta-analysis of COVID-19 generation times and the growth rate r extracted from the SCLAIV+D model output data. The mean plus/minus the standard deviation of R_0 across our CSAs is 4.5 ± 1.8.

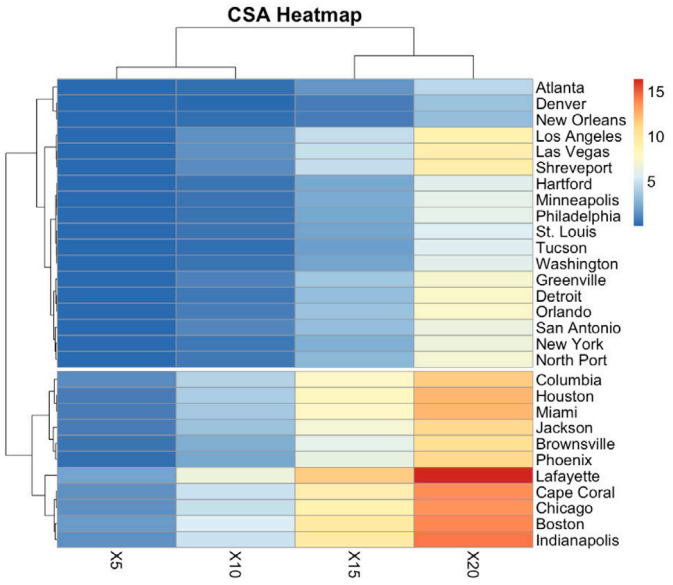

Fig. 4. Heatmap and clusters of CSAs considering the area under the c_{flatten} index curve. The variable X_i represents the area under the curve after i days, starting from the last day with the index equal to 0.

7. Curve flatting index

With the basic set of epidemic parameters fitted to the exponential phase, a second set of behavioural parameters can be fitted to each CSA's incidence data for a couple of weeks beyond the exponential phase. For clarity, we reemphasize that we used the first 15 days of incidence data to fit the exponential phase and the first 30 days to fit the behavioural response parameters, with the best-fitting exponential phase parameters in place for this second fit. The resulting simulations can be used to assess the degree to which the susceptible individuals in susceptible class S had transferred to the social-distancing susceptible class S_r, where we assume that individuals through transmission-reducing or social-distancing behaviour have reduced their risk of infection by an order of magnitude (i.e., by 90%). We then calculated a curve flatting index $c_{\text{flatten}}(t)$ (Getz et al., 2021),

$$c_{\text{flatten}}(t) = \frac{S_r(t)}{S(t) + S_r(t)} \tag{2}$$

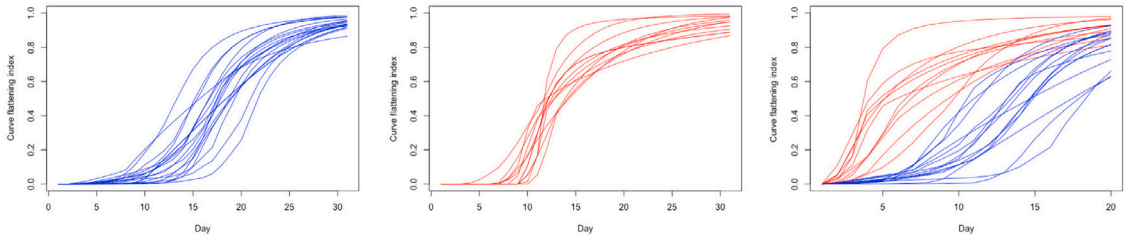

Fig. 5. Curve flatting index: weak-response group (blue, left) and strong-response group (red, centre) over the 30-day range, with the considered 20-day range (right) for the area calculation.

We note that the actual assumption about the value of the reduction in risk (i.e., by 90%) is not critical because the proportion of susceptibles in the response class S_r will adjust dynamically to reflect the real flattening effect when the model is fitted to the data: the index will become larger or smaller depending on whether we make the risk reduction factor less or more stringent respectively. Thus comparisons are always relative across different areas rather than absolute.

We regard the state-derived index expressed in Eq. (2) as a better way to compare curve flattening behaviour than directly comparing the curve-flattening process parameter values themselves, because collinearity among the curve flattening parameters makes it hard to generate an appropriate comparison. We also emphasise that the importance of the SCLAIV+D model fit of social distancing is that it allows us to extract information regarding the social distancing response without having to collect empirical observations. Such empirical data are difficult to collect without the use of GPS or mobile phone technology (but see Grantz et al. 2020), especially for large-scale studies.

At the end of the second fitting procedure, we selected the best 10% of the total resulting fits by looking at the error fit and evaluated mean and standard deviation of the corresponding curve flatting index. The full details of all the fits (e.g. errors associated with the fits, histograms of parameter values associated with the fits), as well as plots of the best fits in each case are reported in SOF3. To obtain a comparative understanding among the CSA responses, we generated a heat map (Fig. 4) as follows. First we captured the temporal evolution of the curve flatting index by calculating the area under the index $c_{\text{flatten}}(t)$ mean curve over time; specifically, over 20 consecutive days, starting from the last day with an index equal to 0 (see SOF1 for details). These values were then used to perform hierarchical clustering (choosing area values at day 5, 10, 15 and 20 to reduce the effect that values on consecutive days are strongly correlated) and classify the CSAs into different groups. We used the R package `pheatmap` and the "complete" method to perform our analysis. This led us to extract two reasonably distinct clusters, as shown in Fig. 4, based on the elbow method for determining the most likely number of distinct clusters in the data (see SOF1 for details). Plots of the curve flatting index values $c_{\text{flatten}}(t)$ for the two different clusters are illustrated in Fig. 5, where we observe that the first group (shown in blue) presents a noticeably weaker response over time compared to the second group (shown in red).

In SOF1 we also present the clusters extracted by considering both area features and the R_0 estimation. Except for the two CSAs (New York-Newark, NY-NJ-CT-PA and Phoenix-Mesa, AZ) which present a particularly high R_0 and therefore are clustered together if we consider three clusters, we observe that the area associated to the curve flattening index tends to be more influential to the clustering process than the R_0 value.

8. Discussion

In this paper, we present an approach to computing R_0 from the initial exponential outbreak phase and extracting an initial social distancing response to this outbreak phase. After these two phases the course of the epidemic depends on the extent to which the initial social distancing response may be relaxed and later reimposed, as we have observed in many of the COVID-19 epidemics in regions and countries around the world (Getz et al., 2021).

To begin, we focused on the initial outbreak phase across the Combined Statistical Areas (CSAs) of the United States. In the literature we find previous work on the estimation of R_0 across the US: in particular, in Sy et al. (2021), the authors estimated R_0 across US counties by assuming that the exponential growth phase began one week prior to the second daily increase in cases and that the period of exponential growth lasted approximately 18 day beyond this start day. They extracted values of R_0 in the range $[0.38, 12.44]$, with median value $R_0 = 1.66$. Our method, on the other hand, searched for the best fitting exponential curve over a two-week window that was allowed to range from the day of the first case (considering a 7-day lagged moving average, see SOF1 for details) for an additional two weeks beyond this first case. Also, we considered population units that were socio-economically more cohesive than counties, because the counties may either include areas not particularly socio-economically connected or they may cut populations into fragments that belong to the same socio-economic unit.

The values we obtained for R_0 spanned an almost 5-fold range of $[1.9, 9.4]$ thereby indicating the precariousness of estimating R_0 from data pertaining to a population that is not closed with respect to the importation versus autochthonous generation of new infections. The mean value we obtained ($R_0 = 4.5 \pm 1.8$) is somewhat higher but still in line with other estimated values of R_0 for COVID-19. In fact, in initial outbreak phases, the occurrence of super-spreading events (Lloyd-Smith et al., 2005) may precipitate the start of an outbreak (Kochańczyk et al., 2020), thereby resulting in higher R_0 estimates than would be obtained if the existence of such superspreading events were taken into account when estimating R_0. In particular, in Kochańczyk et al. (2020), the R_0 estimate across various countries, observed in the exponential phases, resulted in the interval $[4.7, 11.4]$. In Ke et al. (2021), the authors extracted an early epidemic growth rate r for the US equal to 0.28/day, which is consistent with our mean result ($r = 0.27$). The authors calculated R_0 according to the equation derived by Wearing et al. (2005b) and estimated its value in the range $[4.7, 7.5]$ with median $R_0 = 5.9$.

In conclusion, we have presented an objective, mathematically formulated method for estimating R_0 within a month or so of the start of an autochthonous outbreak. Thus our method is more rigorous than previous methods, which appear to be more ad-hoc and do not try to account for the start of the autochthonous outbreak. Specifically, our method involves mathematically fitting the best exponential growth curve in a way that is likely to minimise the impacts of non-autochthonous cases; other methods have no way of eliminating early non-autochthonous cases that do not immediately seed an autochthonous outbreak that gives rise to the exponential phase. In addition, we stress that our approach allows us to compute an initial social distancing response in a population experiencing an epidemic outbreak without the need to have data that can be used to directly assess actual social distancing behaviour. Notably, more than two years into the COVID-19 epidemic, a reliable way beyond our presentation here has not be developed to quantifying social distancing. During this period, however, GPS based mobility data became widely available

with multiple studies finding a relationship between mobility and the early phases of the pandemic (Yabe et al., 2022; Serafino et al., 2022). Thus, the time is now ripe for research to be undertaken to connect mobility data obtained during the Covid-19 pandemic with model-based indicators of population behaviour. Finally, our method applies equally well to an outbreak of any directly transmitted disease or disease process that has an initial exponential growth phase.

Funding

This research was funded by National Science Foundation (NSF) Grant 2032264 (PI: WMG).

CRediT authorship contribution statement

Ludovica Luisa Vissat: Methodology, Formal analysis, Investigation, Visualization, Writing, Editing. **Nir Horvitz:** Methodology, Investigation, Data curation, Visualization, Writing. **Rachael V. Phillips:** Formal analysis, Investigation. **Zhongqi Miao:** Investigation. **Whitney Mgbara:** Investigation. **Yue You:** Investigation. **Alan E. Hubbard:** Methodology, Editing, Supervision. **Wayne M. Getz:** Conceptualization, Methodology, Visualization, Writing, Editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

Wayne M. Getz and Richard Salter are two of three owners of Numerus Inc., which is responsible for developing and maintaining the SCLAIV+D web app used in our exponential curving fitting and social-behaviour-driven curve-flattening analyses.

Appendix A. Supplementary data

Details on methods and analysis presented in the paper can be found in the Supplementary File Online 1 (SOF1: https://ludovicalv.github.io/PDFs/SOF1_CSA.pdf). The guide to run the automatic fitting procedure is presented in the Supplementary File Online 2 (SOF2: https://ludovicalv.github.io/PDFs/SOF2_CSA.pdf) while the output of the fitting procedure for each CSA can be found in the Supplementary File Online 3 (SOF3: https://ludovicalv.github.io/PDFs/SOF3_CSA.pdf). The R code and the data associated with this manuscript can be downloaded from https://ludovicalv.github.io/CSA/.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.epidem.2022.100640.

References

Ali, M.J., Hanif, M., Haider, M.A., Ahmed, M.U., Sundas, F., Hirani, A., Khan, I.A., Anis, K., Karim, A.H., 2020a. Treatment options for COVID-19: A review. Front. Med. 7, 480. http://dx.doi.org/10.3389/fmed.2020.00480.

Ali, S.T., Wang, L., Lau, E.H.Y., Xu, X.-K., Du, Z., Wu, Y., Leung, G.M., Cowling, B.J., 2020b. Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. Science 369 (6507), 1106–1109. http://dx.doi.org/10.1126/science.abc9004.

Anderson, R.M., Donnelly, C., Hollingsworth, D., Keeling, M., Vegvari, C., Baggaley, R., Maddren, R., 2020. Reproduction number (R) and growth rate (r) of the COVID-19 epidemic in the UK: methods of estimation, data sources, causes of heterogeneity, and use as a guide in policy formulation. R. Soc. 2020.

Anderson, R.M., Fraser, C., Ghani, A.C., Donnelly, C.A., Riley, S., Ferguson, N.M., Leung, G.M., Lam, T.H., Hedley, A.J., 2004. Epidemiology, transmission dynamics and control of SARS: the 2002–2003 epidemic. Philos. Trans. R. Soc. London [Biol] 359 (1447), 1091–1105. http://dx.doi.org/10.1098/rstb.2004.1490.

Anderson, R.M., May, R.M., 1992. Infectious Diseases of Humans: Dynamics and Control. Oxford University Press.

Anglemyer, A., Moore, T.H., Parker, L., Chambers, T., Grady, A., Chiu, K., Parry, M., Wilczynska, M., Flemyng, E., Bero, L., 2020. Digital contact tracing technologies in epidemics: a rapid review. Cochrane Database Syst. Rev. (8), http://dx.doi.org/10.1002/14651858.CD013699.

Billah, M.A., Miah, M.M., Khan, M.N., 2020. Reproductive number of coronavirus: A systematic review and meta-analysis based on global level evidence. PloS ONE 15 (11), e0242128. http://dx.doi.org/10.1371/journal.pone.0242128.

Blumberg, S., Lloyd-Smith, J.O., 2013. Inference of R_0 and transmission heterogeneity from the size distribution of stuttering chains. PLoS Comput. Biol. 9 (5), e1002993. http://dx.doi.org/10.1371/journal.pcbi.1002993.

Briz-Redón, Á., Serrano-Aroca, Á., 2020. The effect of climate on the spread of the COVID-19 pandemic: A review of findings, and statistical and modelling techniques. Progr. Phys. Geogr.: Earth Environ. 44 (5), 591–604. http://dx.doi.org/10.1177/0309133320946302.

Chen, J.T., Krieger, N., 2021. 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. 27 (1), S43–S56. http://dx.doi.org/10.1097/phh.0000000000001263.

Davies, N.G., Kucharski, A.J., Eggo, R.M., Gimma, A., Edmunds, W.J., Jombart, T., O'Reilly, K., Endo, A., Hellewell, J., Nightingale, E.S., et al., 2020. Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study. Lancet Public Health 5 (7), e375–e385. http://dx.doi.org/10.1016/S2468-2667(20)30133-X.

Delamater, P.L., Street, E.J., Leslie, T.F., Yang, Y.T., Jacobsen, K.H., 2019. Complexity of the basic reproduction number (R_0). Emerg. Infect. Diseases 25 (1), 1. http://dx.doi.org/10.3201/eid2501.171901.

Dziak, J.J., Nahum-Shani, I., Collins, L.M., 2012. Multilevel factorial experiments for developing behavioral interventions: power, sample size, and resource considerations. Psychol. Methods 17 (2), 153. http://dx.doi.org/10.1037/a0026972.

Getz, W.M., Salter, R., Luisa Vissat, L., Horvitz, N., 2021. A versatile web app for identifying the drivers of COVID-19 epidemics. J. Trans. Med. 19 (1), 1–20. http://dx.doi.org/10.1186/s12967-021-02736-2.

Grantz, K.H., Meredith, H.R., Cummings, D.A., Metcalf, C.J.E., Grenfell, B.T., Giles, J.R., Mehta, S., Solomon, S., Labrique, A., Kishore, N., et al., 2020. The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. Nature Commun. 11 (1), 1–8. http://dx.doi.org/10.1038/s41467-020-18190-5.

Ke, R., Romero-Severson, E., Sanche, S., Hengartner, N., 2021. Estimating the reproductive number R_0 of SARS-CoV-2 in the United States and eight European countries and implications for vaccination. J. Theoret. Biol. (ISSN: 0022-5193) 517, 110621. http://dx.doi.org/10.1016/j.jtbi.2021.110621.

Keeling, M.J., Bjørnstad, O.N., Grenfell, B.T., 2004. Metapopulation dynamics of infectious diseases. In: Ecology, Genetics and Evolution of Metapopulations. Elsevier, pp. 415–445. http://dx.doi.org/10.1016/B978-012323448-3/50019-2.

Khataee, H., Scheuring, I., Czirok, A., Neufeld, Z., 2021. Effects of social distancing on the spreading of COVID-19 inferred from mobile phone data. Sci. Rep. 11 (1), 1–9. http://dx.doi.org/10.1038/s41598-021-81308-2.

Kishore, N., Taylor, A.R., Jacob, P.E., Vembar, N., Cohen, T., Buckee, C.O., Menzies, N.A., 2022. Evaluating the reliability of mobility metrics from aggregated mobile phone data as proxies for SARS-CoV-2 transmission in the USA: a population-based study. Lancet Digital Health 4 (1), e27–e36.

Kochańczyk, M., Grabowski, F., Lipniacki, T., 2020. Super-spreading events initiated the exponential growth phase of COVID-19 with R_0 higher than initially estimated. R. Soc. Open Sci. 7 (9), 200786. http://dx.doi.org/10.1098/rsos.200786.

Kraemer, M.U., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D.M., du Plessis, L., Faria, N.R., Li, R., et al., Open COVID-19 Data Working Group, 2020. The effect of human mobility and control measures on the COVID-19 epidemic in China. Science 368 (6490), 493–497. http://dx.doi.org/10.1126/science.abb4218.

Lehtinen, S., Ashcroft, P., Bonhoeffer, S., 2021. On the relationship between serial interval, infectiousness profile and generation time. J. R. Soc. Interface 18 (174), 20200756. http://dx.doi.org/10.1098/rsif.2020.0756.

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K.S., Lau, E.H., Wong, J.Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., Tu, W., Chen, C., Jin, L., Yang, R., Wang, Q., Zhou, S., Wang, R., Liu, H., Luo, Y., Liu, Y., Shao, G., Li, H., Tao, Z., Yang, Y., Deng, Z., Liu, B., Ma, Z., Zhang, Y., Shi, G., Lam, T.T., Wu, J.T., Gao, G.F., Cowling, B.J., Yang, B., Leung, G.M., Feng, Z., 2020. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N. Engl. J. Med. 382 (13), 1199–1207. http://dx.doi.org/10.1056/NEJMoa2001316.

Liu, Y., Gayle, A.A., Wilder-Smith, A., Rocklöv, J., 2020. The reproductive number of COVID-19 is higher compared to SARS coronavirus. J. Travel Med. http://dx.doi.org/10.1093/jtm/taaa021.

Liu, Y., Rocklöv, J., 2021. The reproductive number of the delta variant of SARS-CoV-2 is far higher compared to the ancestral SARS-CoV-2 virus. J. Travel Med. http://dx.doi.org/10.1093/jtm/taab124.

Lloyd-Smith, J.O., Schreiber, S.J., Kopp, P.E., Getz, W.M., 2005. Superspreading and the effect of individual variation on disease emergence. Nature 438 (7066), 355–359. http://dx.doi.org/10.1038/nature04153.

Management and Budget Office, 2010. 2010 Standards for delineating metropolitan and micropolitan statistical areas. 75 FR 37245.

McCoy, D., Mgbara, W., Horvitz, N., Getz, W.M., Hubbard, A., 2021. Ensemble machine learning of factors influencing COVID-19 across US counties. Sci. Rep. 11 (1), 1–14. http://dx.doi.org/10.1038/s41598-021-90827-x.

Mehmood, K., Bao, Y., Abrar, M.M., Petropoulos, G.P., Saifullah, Soban, A., Saud, S., Khan, Z.A., Khan, S.M., Fahad, S., 2021. Spatiotemporal variability of COVID-19 pandemic in relation to air pollution, climate and socioeconomic factors in Pakistan. Chemosphere (ISSN: 0045-6535) 271, 129584. http://dx.doi.org/10.1016/j.chemosphere.2021.129584.

Ottensmann, J.R., 2017. On the choice of Combined Statistical Areas. Available at SSRN 2955456.

Petersen, E., Koopmans, M., Go, U., Hamer, D.H., Petrosillo, N., Castelli, F., Storgaard, M., Al Khalili, S., Simonsen, L., 2020. Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics. Lancet Infect. Dis. 20 (9), e238–e244.

Rai, B., Shukla, A., Dwivedi, L.K., 2020. Estimates of serial interval for COVID-19: A systematic review and meta-analysis. Clin. Epidemiology Glob. Health http://dx.doi.org/10.1016/j.cegh.2020.08.007.

Rozenfeld, Y., Beam, J., Maier, H., Haggerson, W., Boudreau, K., Carlson, J., Medows, R., 2020. A model of disparities: risk factors associated with COVID-19 infection. Int. J. Equity Health 19 (1), 1–10. http://dx.doi.org/10.1186/s12939-020-01242-z.

Sanche, S., Lin, Y.T., Xu, C., Romero-Severson, E., Hengartner, N., Ke, R., 2020. High contagiousness and rapid spread of severe acute respiratory syndrome coronavirus 2. Emerg. Infect. Diseases 26 (7), 1470.

Scherer, A., McLean, A., 2002. Mathematical models of vaccination. Br. Med. Bull. 62 (1), 187–199. http://dx.doi.org/10.1093/bmb/62.1.187.

Serafino, M., Monteiro, H.S., Luo, S., Reis, S.D., Igual, C., Lima Neto, A.S., Travizano, M., Andrade Jr., J.S., Makse, H.A., 2022. Digital contact tracing and network theory to stop the spread of COVID-19 using big-data on human mobility geolocalization. PLoS Comput. Biol. 18 (4), e1009865.

Siedner, M.J., Harling, G., Reynolds, Z., Gilbert, R.F., Haneuse, S., Venkataramani, A.S., Tsai, A.C., 2020. Social distancing to slow the US COVID-19 epidemic: Longitudinal pretest-posttest comparison group study. PLOS Med. 17 (8), 1–12. http://dx.doi.org/10.1371/journal.pmed.1003244.

Sy, K.T.L., White, L.F., Nichols, B.E., 2021. Population density and basic reproductive number of COVID-19 across United States counties. PloS ONE 16 (4), e0249271. http://dx.doi.org/10.1371/journal.pone.0249271.

Torrats-Espinosa, G., 2021. Using machine learning to estimate the effect of racial segregation on COVID-19 mortality in the United States. Proc. Natl. Acad. Sci. 118 (7), http://dx.doi.org/10.1073/pnas.2015577118.

Travaglio, M., Yu, Y., Popovic, R., Selley, L., Leal, N.S., Martins, L.M., 2021. Links between air pollution and COVID-19 in England. Environ. Pollut. (ISSN: 0269-7491) 268, 115859. http://dx.doi.org/10.1016/j.envpol.2020.115859.

Tsang, H.F., Chan, L.W.C., Cho, W.C.S., Yu, A.C.S., Yim, A.K.Y., Chan, A.K.C., Ng, L.P.W., Wong, Y.K.E., Pei, X.M., Li, M.J.W., et al., 2021. An update on COVID-19 pandemic: the epidemiology, pathogenesis, prevention and treatment strategies. Expert Rev. Anti-Infective Therapy 19 (7), 877–888. http://dx.doi.org/10.1080/14787210.2021.1863146.

Voinsky, I., Baristaite, G., Gurwitz, D., 2020. Effects of age and sex on recovery from COVID-19: Analysis of 5769 Israeli patients. J. Infect. 81 (2), e102–e103. http://dx.doi.org/10.1016/j.jinf.2020.05.026.

Wang, J., Li, W., Yang, B., Cheng, X., Tian, Z., Guo, H., 2020. Impact of hydrological factors on the dynamic of COVID-19 epidemic: A multi-region study in China. Environ. Res. (ISSN: 0013-9351) 110474. http://dx.doi.org/10.1016/j.envres.2020.110474.

Wearing, H.J., Rohani, P., Keeling, M.J., 2005a. Appropriate models for the management of infectious diseases. PLOS Med. 2 (7), http://dx.doi.org/10.1371/journal.pmed.0020174.

Wearing, H.J., Rohani, P., Keeling, M.J., 2005b. Appropriate models for the management of infectious diseases. PLoS Med. 2 (7), e174. http://dx.doi.org/10.1371/journal.pmed.0020174.

Weill, J.A., Stigler, M., Deschenes, O., Springborn, M.R., 2020. Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. Proc. Natl. Acad. Sci. (ISSN: 0027-8424) 117 (33), 19658–19660. http://dx.doi.org/10.1073/pnas.2009412117.

Wu, J.T., Leung, K., Leung, G.M., 2020. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 395 (10225), 689–697.

Yabe, T., Tsubouchi, K., Sekimoto, Y., Ukkusuri, S.V., 2022. Early warning of COVID-19 hotspots using human mobility and web search query data. Comput. Environ. Urban Syst. 92, 101747.

Zhang, H., Liu, Y., Chen, F., Mi, B., Zeng, L., Pei, L., 2021. The effect of sociodemographic factors on COVID-19 incidence of 342 cities in China: a geographically weighted regression model analysis. BMC Infect. Dis. 21 (1), 1–8. http://dx.doi.org/10.1186/s12879-021-06128-1.