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.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Mapping the geodemographics of racial, economic, health, and COVID-19 deaths inequalities in the conterminous US

George Grekousis\textsuperscript{a,b,c}, Ruoyu Wang\textsuperscript{d}, Ye Liu\textsuperscript{a,b,c,*}

\textsuperscript{a} School of Geography and Planning, Sun Yat-sen University, China
\textsuperscript{b} Guangdong Key Laboratory for Urbanisation and Geo-simulation, Sun Yat-sen University, China
\textsuperscript{c} Guangdong Provincial Engineering Research Center for Public Security and Disaster, China
\textsuperscript{d} Institute of Geography, School of GeoSciences, University of Edinburgh, UK

\textbf{ARTICLE INFO}

Keywords: COVID-19 deaths Geodemographics Local fuzzy geographically-weighted clustering USA

\textbf{ABSTRACT}

A large number of studies have examined individual-level factors that increase COVID-19 fatalities. However, no research has focused on the geodemographic classification of the most susceptible communities to COVID-19. In this cross-sectional ecological study, we used local fuzzy geographically-weighted clustering to create the socioeconomic profile of the US counties in relation to COVID-19 death rates. We demonstrate that living in a county which has households with lower income, people with a lack of health insurance, a high African-American percentage, and lower education level, lead to 27.12\% higher COVID-19 death rates than the national median, and 72.56\% higher compared to the least vulnerable counties. Compared to counties with a high COVID-19 death rate, counties with a low COVID-19 death rate have 44.90\% higher annual median household income and nearly double house worth (89.51\% more). Results show that the effects of the COVID-19 pandemic are not universal and that the minoritised and impoverished populations suffer more. Our analysis can effectively pinpoint the most vulnerable counties and importantly allows for understanding the socioeconomic context in which tailored interventions can be applied to mitigate COVID-19 deaths.

1. Introduction

Death is unavoidable and no one can escape from it. But people will not die from the same cause. In fact, the place in which we live and the life we lead have a tremendous effect on our health quality and life expectancy (CDC 2020). Assessing who and where are more at risk from dying from COVID-19 is of great interest to epidemiologists, governments, and policymakers - not only because the pandemic is still developing in many parts of the world but mainly for efficiently reducing mortality in the upcoming years.

The SARS-CoV-2 is here to stay, so we should minimise its long-term repercussions. With widespread vaccination, the virus is expected to cause fewer fatalities. Even so, it is crucial to understand the different sources of COVID-19 transmission and the societal risk factors that accelerate its diffusion to spot those who are more prone to die from COVID-19 (Coccia, 2021a). This will increase the preparedness of countries to cope with COVID-19 by informing health care specialists and policymakers act proactively and design successful strategies that support best practices of COVID-19 prevention (Tavares et al., 2020; Coccia, 2021b).

In particular, the pandemic’s diffusion patterns are due to various environmental, population mobility, and socioeconomic factors (Coccia, 2020a). In regard with environmental factors, manifold studies analyse potential relations between high levels of air pollution, low wind speed, and diffusion of COVID-19, resulting in higher numbers of deaths (Coccia, 2020b, 2021c). For example, Coccia (2021d) showed that the diffusion of COVID-19 in cities with more than 100 days per year exceeding PM10 or ozone limits, a low average speed of the wind, and a lower average temperature, lead to higher numbers of COVID-19 infections and deaths.

Population mobility highly contributes to COVID-19 diffusion (Tieskens et al., 2021). For this reason, various non-pharmaceutical interventions, such as cancelling public events, closing schools, restaurants and workplaces, social distancing, and strict lockdowns, were adopted worldwide (Coccia, 2021e). These interventions aimed to slow down COVID-19 transmission by restricting mobility to prevent the health care system from being overwhelmed (Askitas et al., 2021).

A growing amount of literature uses the weathering hypothesis to

\begin{itemize}
  \item Corresponding author. Sun Yat-Sen University, School of Geography and Planning, Department of Urban and Regional Planning, China.
  \item E-mail addresses: geograik@gmail.com (G. Grekousis), R.Wang-S4@sms.ed.ac.uk (R. Wang), liuye25@mail.sysu.edu.cn (Y. Liu).
\end{itemize}

https://doi.org/10.1016/j.apgeog.2021.102558

Received 14 March 2021; Received in revised form 16 August 2021; Accepted 30 August 2021
Available online 4 September 2021
0143-6228/© 2021 Elsevier Ltd. All rights reserved.
link social inequality to COVID-19 deaths and trace the more vulnerable populations (Adhiikari et al., 2020, Ahmed et al., 2021, Alkire et al., 2020, Azar et al., 2020, Escobar et al., 2021, Grekousis & Liu, 2021, Rodriguez et al., 2020, Selden and Berdahl, 2020, Tai et al., 2021, Tavares et al., 2021). The weathering hypothesis states that chronic exposure to social and economic disadvantage leads to a decline in physical health outcomes (Forde et al., 2019). The mortality of COVID-19 depends on the social determinants of health, such as age, race, occupation, poverty, air quality, lack of education, or unequal access to health facilities (Quan et al., 2021). For example, Hispanic and Black population (Saini et al., 2021; Pan et al., 2021), high population density (Desmet et al., 2020), median household income (Mollalo et al., 2020), low education level, poverty level (Hawkins et al., 2020), number of intensive unit care beds (Andersen et al., 2021) and air quality (Coccia, 2020c) are associated with higher COVID-19 fatalities. A large number of studies adopt mainly regression analysis and spatial regression modelling to understand the structural drivers behind COVID-19 deaths variances and the observed socioeconomic disparities. However, understanding inequalities in the context of COVID-19 requires not only modelling studies that identify the socioeconomic drivers of COVID-19 deaths, but also the geodemographic characteristics of the most susceptible communities. Such understanding is essential for analysing the complex interplay of space, demographics, socioeconomic and health factors on COVID-19 mortality.

In essence, the core research and output focus of geodemographic analysis is the clustering of people based on their demographical, social and economic characteristics, adjusted to the place they live. Geodemographic analysis explores over-, or under-represented demographic, consumption, behavioural, or health patterns across the various neighbourhood types, based on the hypothesis that people residing close together are more likely to be similar than those living further away (Grekousis and Fotis, 2012; Singleton et al., 2016; Long et al., 2019, p. 2019, Grubesic and Durbin, 2020). Thereupon, policies applied can be more beneficial if targeted to specific subgroups instead of a general implementation at a more regional scale. For example, policies against epidemic spread, smoking, chronic lack of nutritious food, education inequality, or unemployment would be more efficient if applied targeted, through geodemographic analysis, to the subdistricts where people residing are more in need (Grekousis et al., 2018; Grekousis, 2018; Xiang et al., 2018).

While there is rich evidence that social determinants (i.e. age, race, poverty, access to health care) and underlying health conditions (i.e. obesity, hypertension, diabetes) are linked to COVID-19 mortality, there is no study that builds the socioeconomic profiling of the communities that are more vulnerable to COVID-19 death (Singu et al., 2020). This study fills this gap and proposes a local-level analysis of COVID-19 through geodemographic segmentation for the conterminous US during the first year of the pandemic (February 6, 2020 to February 5, 2021). Instead of building predictive models, or calculating vulnerability indices, as most previous studies do, we follow a different approach. We conduct geodemographic segmentation (through Local Fuzzy Geographically Weighted Clustering method), which groups US counties based on their health, socioeconomic, and demographic compositions. Counties with similar traits are grouped together and share a common profile, while counties with divergent characteristics are separated. The research assumption is that socioeconomic and health inequalities may exist in the context of the COVID-19 burden. For this reason, we statistically test (through the Mann-Whitney U test) whether community structure may underlie county-level variations of COVID-19 mortality across the US.

Although the proposed approach does not produce cause and effect relationships (as in regression analysis), it offers something equally significant from the policy perspective: a comprehensive profiling of those areas that are more susceptible to high COVID-19 deaths. Furthermore, such analysis allows for statistical comparisons across clusters. In other words, not only a detailed description of each group is provided, but a comprehensive comparison among groups is carried out to assess inequalities across a wide variety of socioeconomic, demographic, and health variables. On the contrary, due to complexity and statistical constraints (i.e. multicollinearity), regression analysis and spatial regression modelling studies tend to finally retain only a small fraction of the original dataset and report results, most of the time for a handful of independent variables (Andersen et al., 2021; Mollalo et al., 2020). Although this offers simplicity, as the focus lies on the most significant social determinants, it neglects the variables that build the socioeconomic context in which people live and act.

Moreover, fuzzy clustering has a significant advantage over other clustering methods. It allows each spatial entity to belong to every cluster with a varying membership value (Grekousis et al., 2011; Betti et al., 2018), avoiding the binary assignment of an object to a single group (when using non-fuzzy clustering methods). In crisp clustering, there is information loss because a binary assignment does not accurately reflect reality. For example, in geodemographic analysis, a spatial entity such as a census tract or a zip code would belong to one, and only one, cluster. In practice, this would mean that the entire population of this spatial entity should be described in a single way (i.e., affluent), which is an exaggeration. For spatial entities having a large population size, as in counties, fuzzy clustering is more detailed than non-fuzzy methods. The reason is that we do not aggregate our results on the basis of only one label, thus avoiding the statistical problem of ecological fallacy, and we delineate a better profile of each spatial entity.

In summary, this study further advances the assessment of the impact of demographic and socioeconomic factors on COVID-19 death by tracing the most disadvantaged population. The provision of which communities are more susceptible to high COVID-19 death rates is essential to inform national, regional, and local strategies. Our analysis can be seen as a tool, supplementary to those existing, to draw better health policies with testing, vaccination, or other non-pharmaceutical interventions. Besides this, it could be used for precise targeting of particularly vulnerable groups, allowing for efficient resource allocation (i.e. medical personnel, vaccines, rapid tests) to geographical areas most in need.

2. Materials and methods

2.1. Sample and data

This cross-sectional ecological study analysed COVID-19 deaths data across 3108 counties in the conterminous US. Data analysis refers to the first year of the pandemic, from February 6, 2020 (day of the first reported case) up to February 5, 2021. The COVID-19 deaths were collected from USAFacts (USAFacts 2021). The number of confirmed COVID-19 deaths is adjusted by the population of each county as derived from the U.S. Census Bureau 2015–2019 American Community Survey (U.S. Census Bureau, 2021). The COVID-19 death rate for each county is calculated as the cumulative deaths per 100,000 inhabitants as of February 5, 2021.

2.2. Measures of variables

Similar to other studies (Andersen et al., 2020; Mollalo et al., 2020), an initial set of 102 variables hypothesised for contributing to potential higher COVID-19 death rates was compiled from various databases across three themes:

- **a)** Demographic (i.e. population, population density, gender, race, age).
- **b)** Socioeconomic (i.e. education, occupation, income, poverty).
- **c)** Underlying health conditions that predispose to COVID-19 (i.e., obesity, diabetes, high blood pressure, heart disease, smoking).

From the original set of 102 variables, we excluded 29 variables that exhibited high multicollinearity. We also merged some variables to
produce fewer and more meaningful categories. Moreover, we excluded those variables so that their cluster centres did not vary significantly across cluster centres. This was done after several clustering trials and tests. Finally, we retained 25 variables of which five are demographic, eight socioeconomic, and 12 related to people’s health condition (Table 1, for variables’ definitions and data sources see Table S1 in Supplementary Material). To partition our dataset, we used the LFGW clustering approach, explained next.

2.3. Model and data analysis procedure

We used the Local Fuzzy Weighted Geographically Weighted Clustering (LFGWC) method to segment the dataset as introduced in Grekousis (2021). This method makes use of the fuzzy c-means algorithm which is the most well-known algorithm in fuzzy clustering (Bezdek, 2013), and has been widely applied in geodemographic applications (Grekousis et al., 2013). However, the fuzzy c-means algorithm does not take into account any geographical aspects, which is an important drawback when segmenting spatial data. The LFGW overcomes the absence of a spatial framework by integrating geography into the fuzzy c-means algorithm. To do so, the LFGW updates the membership value of each spatial entity for the assigned cluster, according to the membership values of the entities within its neighbourhood. Spatial weights are created using either a spatial interaction model or distance decay functions.

The objective function to be minimised in LFGWC is (Eq. (1)):

$$J = \sum_{i=1}^{n} \sum_{k=1}^{c} u_{ik}^\alpha \left(\frac{d(x_i, v_k)^2}{\sum_{j=1}^{n} w_{ij}^\beta} \right)^{1/\beta}$$

where:
- $U$ is the membership matrix.
- $V$ is the clusters’ centre matrix.
- $X$ is the data matrix.
- $c$ is the number of clusters.
- $n$ is the number of spatial objects.
- $u_{ik}$ is the membership value of the $k$-th object to the $i$-th cluster.
- $m$ is the fuzziness (a weighting exponent).
- $\eta_i$ is the cluster centre of the $i$-th cluster.
- $x_k$ is the data point $k$, and;
- $*$ is any norm expressing the similarity between a data point and the cluster centre.

In contrast to fuzzy c-means, which does not integrate space, LFGWC modifies the membership matrix values in each iteration by using Eq. 2

$$\eta_i^t = \eta_i^{t-1} + \beta \sum_{k=1}^{c} u_{ik}^{\alpha} \left(\frac{d(x_i, v_k)^2}{\sum_{j=1}^{n} W_{ij}^\beta} \right)^{1/\beta}$$

where:
- $W_{ij}^t$ is the $n \times n$ spatial weights matrix, calculated using the inverse distance decay function. Spatial weights are row-standardised for scaling in the range between 0 and 1, so that the number of neighbours has no effect on the final results.
- $d$ is the distance between spatial objects areas $i$ and $j$.

| Variables | Cluster centres | Median values | Global statistics |
|-----------|----------------|---------------|------------------|
|           | Group 1 (High) | Group 2 (Moderate) | Group 3 (Low) | Q1 | Median | Q3 | Mean |
| Population density (people/km²) | 64.61 | 100.69 | 123.22 | 16.31 | 19.49 | 17.51 | 6.70 | 17.43 | 45.75 | 105.73 |
| Africa (%) | 15.66 | 7.38 | 4.82 | 7.73 | 2.20 | 1.23 | 0.70 | 2.33 | 10.33 | 9.16 |
| Disabled (%) | 18.16 | 15.29 | 13.52 | 18.36 | 15.03 | 12.44 | 12.60 | 15.05 | 18.09 | 15.55 |
| House median value (US $) | 113331.81 | 150284.95 | 189883.55 | 98200.00 | 132850.00 | 186100.00 | 97600.00 | 127200.00 | 173900.00 | 152685.19 |
| Rent (median) (US $) | 681.88 | 770.54 | 844.47 | 650.00 | 728.00 | 826.00 | 632.00 | 714.00 | 844.00 | 770.82 |
| Bachelor and higher (%) | 38.57 | 35.65 | 32.86 | 36.10 | 33.25 | 30.40 | 31.00 | 33.50 | 36.70 | 34.05 |
| Disabled (%) | 18.16 | 15.29 | 13.52 | 18.36 | 15.03 | 12.44 | 12.60 | 15.05 | 18.09 | 15.55 |
| Median income (US $) | 44489.64 | 53880.25 | 60766.87 | 42216.00 | 52357.50 | 61173.00 | 44123.50 | 51658.00 | 59585.00 | 53305.43 |
| No health insurance (%) | 18.70 | 16.45 | 14.30 | 18.30 | 15.80 | 12.50 | 12.05 | 15.10 | 19.20 | 16.57 |
| Heart disease mortality (per 100,000 population) | 86.98 | 75.85 | 69.18 | 87.70 | 76.15 | 66.40 | 66.20 | 75.80 | 86.00 | 77.05 |
| Pulmonary disease (%) | 9.36 | 7.61 | 6.45 | 9.50 | 7.70 | 6.00 | 6.30 | 7.60 | 9.00 | 7.76 |
| Diabetes (%) | 12.60 | 10.59 | 9.40 | 12.80 | 10.50 | 8.70 | 9.10 | 10.50 | 12.20 | 10.85 |
| High blood pressure (%) | 36.36 | 31.82 | 29.27 | 37.10 | 31.55 | 28.20 | 28.90 | 31.60 | 35.50 | 32.39 |
| Obesity (%) | 37.80 | 35.02 | 32.97 | 38.60 | 35.40 | 32.20 | 32.70 | 35.60 | 38.10 | 35.21 |
| Sleep less than 7 h (%) | 39.70 | 36.60 | 34.26 | 40.30 | 36.70 | 33.00 | 33.80 | 36.80 | 39.70 | 36.83 |
| No physical activity (%) | 31.45 | 26.48 | 23.42 | 32.20 | 26.60 | 22.30 | 23.30 | 26.50 | 30.65 | 27.01 |
| Smoking (%) | 24.44 | 21.06 | 18.77 | 25.20 | 21.35 | 18.10 | 18.60 | 21.10 | 24.00 | 21.31 |
| Physically unhealthy days (days per month) | 15.79 | 13.37 | 11.86 | 16.10 | 13.40 | 11.10 | 11.70 | 13.50 | 15.40 | 13.63 |
| Mentally unhealthy days (days per month) | 17.10 | 14.95 | 13.47 | 17.40 | 15.20 | 12.80 | 13.40 | 15.10 | 16.70 | 15.12 |
| PM2.5 | 8.41 | 7.64 | 6.92 | 8.70 | 7.90 | 6.70 | 6.40 | 7.90 | 8.80 | 7.64 |
| Covid-19 deaths (per 100,000 population) | 179.52 | 151.75 | 132.59 | 176.98 | 138.78 | 102.56 | 82.57 | 139.12 | 205.58 | 155.14 |

Note: High = group with higher cluster centres than the 75th quartile for most of the variables; Moderate = group with cluster centres that fluctuate around the median values; Low = group with cluster centres that lie below the median values for the majority of the variables. Covid-19 data refer to the conterminous US (county level) for the first year of the pandemic (2020/2/6–2021/2/5).
$d_{ab}$ is a distance value after which weights are zero.

$a, b$ are user-defined parameters to control the distance decay and the weight of the populations, respectively.

The complete mathematical formation of the algorithm is presented in detail by Grekousis (2021) and is not repeated here. To evaluate the clustering results, we applied the Partition Coefficient (PC), Classification Entropy (CE), the Partition Index (SC), and the IFV validation indices (Bensaid et al., 1996; Hu et al., 2008). We also used these validation indices to assess the appropriate number of clusters that the dataset would be segmented.

To assess if the variables’ distributions are statistically significant different across clusters, we used the Mann–Whitney U test (Mann et al., 1947). The null hypothesis to be tested is that clusters with high COVID-19 death rates have similar characteristics to clusters with low or average COVID-19 death rates. For example, the null hypothesis would be that a cluster having high COVID-19 death rates, also has similar annual household income, racial composition, or underlying health conditions with clusters having low or around the median COVID-19 death rates. Three significance levels for $\alpha$ (0.05, 0.01, and 0.001) were selected and if the p-values resulting from the test were to be less than $\alpha$, then there would be compelling evidence to reject the hypothesis and conclude that the distributions compared are statistically significant different.

For the above example, this would conclude that clusters with high COVID-19 death rates do not share common characteristics (at least for some variables) with clusters with low or average death rates. This would also lead to the identification of the demographic, socioeconomic and health characteristics that are linked to higher COVID-19 death rates and the subsequent tracing of the communities that are more susceptible to COVID-19 death.

Additionally, to identify any regional patterns we analysed how the clusters are distributed across the four US regions (Northeast, Midwest, South, East) as defined by the US Census Bureau (U.S. Census Bureau 2019). Lastly, we studied how clusters are distributed across the Metropolitan Statistical Areas (MSAs) and the non-Metropolitan Statistical Areas (non-MSAs). The designation purpose of MSAs and non-MSAs does not define an urban-rural classification (U.S. OMB, 2010). MSAs signify an extensive area with a core having substantial population (over 50,000), and neighbouring communities that interact highly both economically and socially with the core (U.S. OMB, 2010). In this respect we use MSAs and non-MSAs to trace whether areas that are linked at a higher degree are more likely to experience higher COVID-19 death rates.

Cluster labelling and pen portraits were then derived. Pen portraits are small descriptive analyses of the clusters that draw upon their main identifiable characteristics. To depict fuzzy clustering results, we used two methods. First, we created a map where each county is assigned to the cluster with the highest membership value. From the policy perspective this is a straight-forward way to communicate the clustering results. Second, we depicted the maximum membership value per county. Counties with high maximum values are typical for their clusters. On the other hand, areas with low maximum values are not that typical for their cluster and share common characteristics with one or more other clusters, something that allows for a deeper analysis. More importantly, by depicting the varying membership values in GIS maps and by applying spatial autocorrelation statistical methods, spatial heterogeneity or spatial dependence can be traced (Grekousis and Giais, 2019).

3. Results

We used the LFGW clustering and we tested various parameter settings. We only report here the settings and the results of the clustering that yielded the more separated and compact clusters, as estimated through PC, CE, SC and IFV validation indices (a more thorough analysis on how we determined the optimal number of clusters is provided in Text S1 and Fig. S1 in Supplementary Material). The following parameters were applied for clustering: number of clusters = 3, weighting scheme: inverse distance squared, distance threshold = 100 km, fuzziness m = 2, variables weighting = 0.8.

The output cluster centres and the global statistics per variable are presented in Table 1 (the global statistics refer to the entire unpartitioned set of observations for the conterminous US; $n = 3108$). To analyse results we plot the standardised values of each variable in a boxplot (Fig. 1). We used standardised values to overcome different units and magnitudes across variables. In this way we can better compare all variables as they are now expressed in z-score values centered to have mean 0 and scaled to have standard deviation 1. To compare the groups, we plotted the standardised cluster centres over the boxplots. The coloured dots refer to the standardised cluster centres of each group per variable. Dots of the same colour belong to the same group/cluster. The further apart the groups’ standardised centres lie, the more the separation among the clusters and the better the clustering scheme.

We observe that the three groups are well separated for all variables. The only exception is for the population density variable which does not differentiate across the three groups. We notice that Group 1 (High) has higher cluster centres than the 75th quartile (Q3) for most of the variables, Group 2 (Moderate) cluster centres fluctuate around the median values, and Group 3 (Low) cluster centres lie below the median values for the majority of the variables (Fig. 1).

To analyse and compare the characteristics of each group we depict the boxplots of each variable per group (Fig. 2, Fig. 3). We also plot the global median value across the boxplots to better compare with the national median values (Fig. 2). To assess whether the variables across groups are statistically significant different we calculated the Mann-Whitney median $U$ test for all combinations of groups (Group 1 to Group 2, Group 2 to Group 3, and Group 1 to Group 3). A p-value is calculated, and if smaller than the significance level $\alpha = 0.001$ we conclude that groups have statistically significant different medians in the variables studied. With the exception of population density, all other variables have statistically significant different medians (p-value $<0.001$) across all group combinations. This means that groups are well separated and that we can use the variables’ median values to delineate them.

In regard to COVID-19 death rates, groups are statistically significant different (p-value $<0.001$) (Fig. 2A). The median value of death rate for Group 1 is 176.98 (deaths per 100,000 people) while the global median is 139.12, a 27.21% difference (Fig. 2A, Table 1). Group 2 has a median value of 138.78 very similar to the national (global) median (139.12). Group 3 has a median value of 102.56, which is 26.28% less than the global median.

Taking COVID-19 death rate as the leading variable to describe each group, we label Group 1: High, in terms of having higher death rates than national (27.21% more), Group 2: Moderate, in terms of having similar death rates with the national rate, and Group 3: Low, as this group lies well below the median national death rate (26.28% less).

To identify if these variances in COVID-19 death rates are related to other socioeconomic and underlying health condition variables, we further examined how these variables vary across the three groups. In Fig. 2B, C, D we present three socioeconomic variables (Annual median household income, Housing value, Rent). Annual median household income is 30.99% lower in High group counties (42,216 US dollars) compared to Low group counties (61,173 US dollars). This potentially indicates that income inequality has an impact on COVID-19 death rates as those with lower income are likely to suffer from higher COVID-19 death rates. An even larger inequality is observed for housing value, with the Low group having houses worth on median values 186,100 US dollars, almost double (89.51% more) the value of houses in the High group counties (98,200 US dollars). On median values, rent is 27.08% higher in Low group (826 US dollars) compared to the High group (650 US dollars).

The rest of the variables are presented in the same way in Fig. 3. The boxplots show that the differences are evident, with all median values
being statistically significant different (p < 0.001) across the groups. We will further discuss these results in the next section.

- High (Group 1)

Counties belonging in High group (n = 1047) (Fig. 4), have on median values 72.56% higher COVID-19 death rate than counties in Low group, a significant difference. The median percentage of African-Americans in this group (7.73%) is 3.5 times higher than the Moderate group (2.20%), and nearly six times higher than the Low group (1.23%). This signifies that counties with more African-Americans are more prone to higher COVID-19 death rates. The education level seems also to be linked to higher COVID-19 death rates. The High group has on median values almost half the share of the population possessing a Bachelors’ or higher degree (14.4%) compared to Low group (27.60%). The High group counties seem to be in economic stress in multiple ways. As shown in Fig. 2B and C, households in this group have on median values 30.99% less annual income than the counties belonging in the Low group, while their housing units are worth almost the half value compared to the houses of the Low group. At the same time the unemployment rate in High group (6.50%) is 75.68% more than the unemployment rate in Low group (3.70%). The above economic disparities may also explain why the share of population living under the poverty level is almost double in High group (19.70%) than the Low group (10.00%). Another important aspect, potentially related to COVID-19 deaths, is that the percentage of population that lacks health insurance is nearly 50% more for the High group (18.30%) than the Low group (12.50%).

In addition, the underlying health conditions of the people living in the High group are worse on median values for all variables used in this study than the Low group counties. Indicatively, for the High group, the percentage of population with the pulmonary disease (9.50%), is 58.33% higher than the High group (6.00%). Similarly, the percentage of population with diabetes (12.80%), is nearly 50% higher than the High group (8.70%). Differences ranging from 20% to 60% between the High and the Low groups are evident in other underlying health conditions variables (i.e. strokes mortality, smoking, physically unhealthy days, mentally unhealthy days). Lastly, on median values, PM2.5 is 29.85% higher in the High group (6.7) than in the Low group (6.7).

Geographically, counties belonging to High group are mostly concentrated on the south statistical region (60.77%) (Fig. 4, Table S2 in Supplementary Material). Another 15.57% of the Midwest counties belong to the High group. Only 23.23% of the non-MSA counties belong to this group, while 40.04% of the non-MSA counties belong to the High group (Table S2 in Supplementary Material).

- Moderate (Group 2)

The Moderate group contains n = 944 counties (Fig. 4). The median values for most of the variables lie close to the national (conterminous US) median values (Fig. 1). The COVID-19 death rate is 138.78 (national median = 139.12), the African-American population share is 2.20% (national median = 2.33%), the median annual household income is 52,357.50 US dollars (national median = 51,658 US dollars), and the unemployment rate is 4.90% (national median = 4.90%) (Table 1). Similar median values are reported for the other variables.

Regionally-wise 39.17% of the Northeast counties and 35.52% of the Midwest counties belong in this group (Fig. 4, and Table S2 in Supplementary Material). There is no difference between the MSA and non-MSA counties for the Moderate group.

- Low (Group 3)

Compared to the Moderate group median values, Low group has 26.09%, COVID-19 death rates, almost half the African-American population share, 21.99% less percentage in population with disability, 40.08% higher housing value, 40.82% higher percentage of population having a Bachelors’ or higher degree, 16.84% higher annual median income, 28.05% less share of population below the poverty level, 17.13% less heart disease morality rate, 22.07% less percentage of population with pulmonary disease, and 15.18% less PM2.5 pollution.
Geographically, counties in the Low group concentrate on the eastern part of the Northeast region (58.53%), and can be found in groups in both the Midwest (48.91%) and West (69.49%) regions (Fig. 4, Table S2 in Supplementary Material). Smaller enclaves of Low group counties are also found in the South region (13.19%) in and around MSAs as for example in San Antonio (Texas), Atlanta (Georgia), or Miami (Florida). In fact, 45.70% of the metropolitan areas belong to the Low group.

In addition, we depict the maximum membership value per county in Fig. 5. Counties with a maximum membership value less than 50% indicate that there is also a second group with a relative high membership value. These counties require further study to avoid assigning to them the general characterisation of the dominant group and instead use a combination of the characteristics of the groups in which they participate at a high level. This is extremely helpful when it comes to policy actions. Policies can be tailored according to each county’s characteristics avoiding the binary assignment as a result of other clustering methods. For example, if an intervention should be implemented to those more vulnerable then we could start with populations in the High group that have the highest membership values, indicating counties more typical to High groups and thus more prone to higher COVID-19 death rates. For counties that may share characteristics from two groups, policy implementations can be also adjusted to reflect this case.

4. Discussion

Robust identification of the most vulnerable communities in COVID-19 deaths can facilitate better and more targeted mitigation and control policies. Based on the weathering hypothesis, which proposes that chronic exposure to social and economic disadvantage may lead to physical health decline (Forde et al., 2019), we study how county-level inequalities on demographic, socioeconomic, and health data are linked to COVID-19 death rates. Through geodemographic analysis, we traced the socioeconomic characteristics of communities with varying COVID-19 death rates. In specific, employing local fuzzy geographically
weighted clustering, we identified the main traits of residents living in counties with high, moderate, or low COVID-19 death rates (Table S3 in Supplementary Material).

We found that the most vulnerable counties are those in the High group with COVID-19 death rates (median value) 27.21% higher than the national median, and 72.56% higher than the Low group. High and Low groups have statistically significant different median values and evident inequalities across all variables. Compared to the national median, the High group counties have on median values 3.3 times higher percentage of African-American population which is consistent with studies showing that COVID-19 mortality disproportionately affects the Black population in the US (Golestaneh et al., 2020; Pan et al., 2020; Prata et al., 2021; Rogers et al., 2020). Moreover, the High group counties have on median values 12.17% higher percentage of people...
with a lower education level, and 24.28% higher percentage of people working in the manufacturing sector. Households in counties of the High group get 22.80% less annual income than the national median, and 30.99% less than Low group. A more striking inequality is that the percentage of the population with a lack of health insurance is nearly 50% higher in the High group than the Low group. This finding is potentially related to the population’s deteriorated health condition in this group compared to the other groups. People living in High group counties have 20%–58% higher population percentages with underlying health conditions than the Low group. This inequality in comorbidities may explain the large difference in COVID-19 death rates compared to the Low group. Air pollution is higher in High group with a nearly 30% higher PM2.5 value. A significant finding is that High group counties are concentrated in the South region, with a 60.77% of the region’s counties being labelled as High. This is consistent with other works who showed that the Southeast is marked by high levels of pre-existing health conditions and socio-ecological vulnerability to COVID-19 (Snyder et al., 2020).

Results also showed that non-MSA counties (areas with lower degree of interaction) have more counties in High group than the MSAs. In fact, it seems that COVID-19 death rates are less where interaction is higher (i.e. 45.70% of MSAs belong to Low group). This finding should not drive us to the wrong conclusion that social and economic interaction lead to lower COVID-19 death rates. On the contrary, COVID-19 is highly contagious and for this reason, social distancing is implemented as a first hinder to its spread. The reason that the non-MSAs counties have higher COVID-19 death rates, although interactions are less, can be partially attributed to the fact that 40.04% of them belong to the High group which has on median values nearly 50% higher percentage of people lacking health insurance than the Low group, and in general, population with deteriorated health (higher values in underlying health conditions) than the Moderate and the Low group. Another reason may be, that in areas with less social and economic interactions, people may get a false sense of security against the virus and relax the social distancing measures.

We also didn’t trace any statistically significant variation in population density across the groups. Although population density has a role in the beginning of the pandemic, as the most densely populated areas faced the first pandemic’s wave first (potential as ports of entry through transportation hubs and due to lack of reaction), in a course of one year the virus spread to less densely populated areas as well, something confirmed in other studies (Snyder et al., 2020). As such, population density may be linked to COVID-19 deaths in the initial waves rather than being a permanent factor that influences pandemic through time.

Similar to population density, we didn’t find statistically significant variations for the population over 60 (as with other non-significant variables we did not graphically present this output). Initially we assumed that counties with larger proportion of their population over 60 would have higher COVID-19 death rates, as this disease is especially deadly to older individuals. However, such a trend was not identified. In fact, as shown in other studies, counties around major cities and those in the Southeast (which tend to have higher COVID-19 mortality) have a slightly lower proportion over 60, while those in the upper Midwest and Northeast have larger over-60 populations (Snyder et al., 2020).

The least vulnerable counties are those in Low group. Compared to the national median, counties in the Low group have 26.28% lower COVID-19 death rates, nearly half percentage of African-Americans, a higher percentage of population with lower incomes and higher unemployment, healthier population (10%–30% lower percentages in health-related variables), and less PM2.5 pollution - something consistent with other works (Coccia, 2021). The counties of this group concentrate on the Northeast and West regions and on MSA counties (45.70% of MSAs belong to High group).

The above results reveal that minoritised and impoverished populations are more vulnerable to COVID-19. These population groups cannot easily adapt to preventive measures such as isolation. Due to socioeconomic status (i.e. having lower income and higher unemployment), and working environment (i.e. working in the manufacturing sector), they are more exposed to infection. Moreover, comorbidities often associated with specific population groups create a complex, dynamic context that urges special attention from governments (Qeadan et al., 2021). For instance, systematic inequalities and discrepancies in socioeconomic status are linked to higher incidence of asthma, hypertension, and obesity, which have been linked in turn to higher risk of severe illness from COVID-19 (Singu et al., 2020). Here, we show that

Fig. 5. Maximum membership value per county. Counties with larger values are more typical for their clusters. Data refer to the conterminous US (county level) for the first year of the pandemic (2020/2/6–2021/2/5).
counties of the High group have up to nearly 60% higher population percentages with comorbidities than the Low group (i.e. pulmonary disease 58.33% higher, diabetes 47.13% higher, heart disease mortality 45.45% higher, high blood pressure 31.56% higher, than the Low group). Because the above conditions exert a disproportionate burden on disadvantaged populations and heighten the risk for individuals being infected with SARS-CoV-2, there is a growing need for governments to effectively allocate resources to the most vulnerable communities.

Geodemographic analysis can successfully contribute to this need, as it allows for better resource allocation and spatial planning by evaluating the underlying characteristics of the population studied (Singleton and Spielman, 2014). This is the first study to analyse COVID-19 deaths under a geodemographic prism to the best of our knowledge. The strength of the present COVID-19 related geodemographic analysis is not only that it assesses who are the most vulnerable by creating the community’s profile, but it also determines where the communities most susceptible to COVID-19 death are located. Using geography, we explore socioeconomic, health, and spatial inequalities to inform public health specialists. Health officials and policymakers can subsequently draw targeted policies that prioritise those areas and subpopulations more in need.

5. Conclusions

The main limitation of this study is that registering a death attributed to COVID-19 may become problematic. The US National Center for Health Statistics (NCHS), within the Centres for Disease Control and Prevention (CDC), which compiles COVID-19 data supplied by the 50 states, continually reminds us that COVID-19 deaths at any given time are undercounted as a result of a lag in reporting deaths (Fineberg, 2020). Another problem is that COVID-19 deaths may be underreported due to a missed COVID-19 death diagnosis (for example, people dying at home and not tested for COVID-19). On the other hand, presumed COVID-19 diagnosis for individuals who actually died from other causes may inflate COVID-19 deaths (Fineberg, 2020). Instead of the confirmed COVID-19 deaths data, excess mortality could be used to assess the pandemic’s total impact on mortality (Roser et al., 2020). However, by the time this research was conducted (February 2021), no official estimation of excess deaths for 2020 existed at the county level for the US. In the absence of excess deaths data at the county level, and similar to others (Andersen et al., 2020; Stokes et al., 2020), we rely on the confirmed COVID-19 deaths registered by each county.

Another limitation is the geographical scale of the analysis. A more detailed scale (i.e. zip code) would offer better insights. However, the finest spatial scale that COVID-19 data is available for the entire US is at the county level to the best of our knowledge. Future research should focus on smaller geographical areas allowing for better assessing the community traits related to increased risk for COVID-19 death. Similar to others (Snyder et al., 2020), we did not take into account the difference between nationwide shelter-in-place or lockdown policies or other political factors. Such policies/factors may positively or negatively affect COVID-19 deaths depending on their duration and the epidemic stage that they were implemented, something that has been already reported at the country level in many studies (Askitas et al., 2021). However, isolating these effects at the county level would be laborious and beyond the scope of this research. To address this limitation, we aggregate results to refer to the end of the first year of the pandemic, a time when most of the citizens were well informed about COVID-19 health threats. Even though non-pharmaceutical interventions varied a lot across states (Mollalo et al., 2020), we anticipate that people practiced social distancing not only because they were ordered but mainly because they were influenced by the public sentiment of protection against the virus. Future research should further study this aspect.

Pandemics have been characterised more as a social problem than a healthcare problem (Singu et al., 2020). Demographical, racial, social, economic, and health disparities exist in the context of the COVID-19 burden (Nana-Sinkam et al., 2021; Poulson et al., 2021). As such, communities are vulnerable for different reasons, and therefore tailor-made interventions should be prioritised based on the specific profile of each area. The use of geodemographic profiles in COVID-19 analysis, as proposed in this study, enhances our understanding of higher COVID-19 fatality when socioeconomic and health inequalities are evident across counties. Policy-wise, the take-home message of this study is that the COVID-19 death rates depend on a complex interplay of demographic, socioeconomic, and health-related factors. For this reason, policymakers and health professionals need to exercise considerable caution when they design public interventions so that they integrate these discrepancies in their policies. Our findings convey and confirm an important message that the most socially vulnerable bear a disproportionate burden on deaths and this should be of interest to those that draw health-related policies worldwide.

Funding

This work was supported by the National Natural Science Foundation of China, grant number [41871140].

Declaration of competing interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apgeog.2021.102558.

References

Adhikari, S., Pantaleo, N. P., Feldman, J. M., Ogedegbe, O., Thorpe, L., & Troxel, A. B. (2020). Assessment of community-level disparities in coronavirus disease 2019 (COVID-19): Infections and deaths in large US metropolitan areas. JAMA Netw. Open, 3(7), Article e2016938.

Ahmed, R., Jamali, O., Ishak, W., Nabi, K., & Mustafa, N. (2021). Racial equity in the fight against COVID-19: A qualitative study examining the importance of collecting race-based data in the Canadian context. Trop. Dis. Travel Med. Vac., 7(1), 15.

Alikri, S., Dirksen, J., Nogales, R., & Oldiges, C. (2020). Multidimensional poverty and COVID-19 risk factors: A rapid overview of interlinked deprivations across 5.7 billion people. OPHI briefing 53, oxford poverty and human development initiative. University of Oxford.

Andersen, L. M., Harden, S. R., Sugg, M. M., Runkle, J. D., & Lundquist, T. E. (2021). Analyzing the spatial determinants of local Covid-19 transmission in the United States. The Science of the Total Environment, 754, 142396.

Askitas, N., Tatsiramos, K., & Verheyden, B. (2021). Estimating worldwide effects of non-pharmaceutical interventions on COVID-19 incidence and population mobility patterns using a multiple-event study. Scientific Reports, 11(1), 1972.

Azar, K. M. J., Shen, Z., Romanelli, R. J., et al. (2020). Disparities in outcomes among COVID-19 patients in a large health care system in California. Health Affairs, 39(7), Article e2016938.

Bensaid, A. M., Hall, L. O., Beadek, J. C., Clarke, L. P., Silbiger, M. L., Arrington, J. A., & Murrath, R. F. (1996). Validity-guided (re) clustering with applications to image segmentation. IEEE Transactions on Fuzzy Systems, 4(2), 112–123.

Betti, G., Gagliardi, F., & Verma, V. (2018). Simplified Jackknife variance estimates for fuzzy measures of multidimensional poverty. International Statistical Review, 86(1), 68–86.

Beadek, J. C. (2013). Pattern recognition with fuzzy objective function algorithms. Springer Science & Business Media.

Centers for Disease Control and Prevention. (2020). Social determinants of health. https://www.cdc.gov/socialdeterminants/index.htm.

Coccia, M. (2020a). An index to quantify environmental risk of exposure to future epidemics of the COVID-19 and similar viral agents: Theory and Practice. Environmental Research, 110135. https://doi.org/10.1016/j.envres.2020.110135

Coccia, M. (2020b). How do low wind speeds and high levels of air pollution support the spread of COVID-19? Atmos. Pollut. Resear., PII. https://doi.org/10.1016/j.apr.2020.10.002. S1309-1042(20)30293-2.

Coccia, M. (2020c). Factors determining the diffusion of COVID-19 and suggested strategy to prevent future accelerated viral infectivity similar to COVID. The Science of the Total Environment. https://doi.org/10.1016/j.scitotenv.2020.138474, 138474, 20-APR-2020.

Coccia, M. (2021a). Preparedness of countries to face covid-19 pandemic crisis: Strategic positioning and underlying structural factors to support strategies of prevention of...
