Time-series clustering for home dwell time during COVID-19: what can we learn from it?

Xiao Huang [1]*, Zhenlong Li [2], Junyu Lu [3], Sicheng Wang [2,4], Hanxue Wei [5], and Baixu Chen [6]

[1] Department of Geosciences, University of Arkansas; xh010@uark.edu
[2] Department of Geography, University of South Carolina; zhenlong@sc.edu
[3] School of Community Resources and Development, Arizona State University; Junyu.Lu@asu.edu
[4] Edward J. Bloustein School of Planning and Public Policy, Rutgers, The State University of New Jersey; sicheng.wang@rutgers.edu
[5] Department of City and Regional Planning, Cornell University; hw652@cornell.edu
[6] Department of Computer Science and Engineering, University of Michigan; chenbaix@umich.edu

* Correspondence: xh010@uark.edu

Abstract: In this study, we investigate the potential driving factors that lead to the disparity in the time-series of home dwell time, aiming to provide fundamental knowledge that benefits policy-making for better mitigation strategies of future pandemics. Taking Metro Atlanta as a study case, we perform a trend-driven analysis by conducting Kmeans time-series clustering using fine-grained home dwell time records from SafeGraph, and further assess the statistical significance of sixteen demographic/socioeconomic variables from five major categories. We find that demographic/socioeconomic variables can explain the disparity in home dwell time in response to the stay-at-home order, which potentially leads to disparate exposures to the risk from the COVID-19. The results further suggest that socially disadvantaged groups are less likely to follow the order to stay at home, pointing out the extensive gaps in the effectiveness of social distancing measures exist between socially disadvantaged groups and others. Our study reveals that the long-standing inequity issue in the U.S. stands in the way of the effective implementation of social distancing measures. Policymakers need to carefully evaluate the inevitable trade-off among different groups, making sure the outcomes of their policies reflect interests of the socially disadvantaged groups.

Keywords: COVID-19; home dwell time; time-series clustering; stay-at-home orders

Highlights:

- We perform a trend-driven analysis by conducting Kmeans time-series clustering using fine-grained home dwell time records from SafeGraph.
- We find that demographic/socioeconomic variables can explain the disparity in home dwell time in response to the stay-at-home order.
- The results suggest that socially disadvantaged groups are less likely to follow the order to stay at home, potentially leading to more exposures to the COVID-19.
- Policymakers need to make sure the outcomes of their policies reflect the interests of the disadvantaged groups.

Introduction

The coronavirus disease 2019 (COVID-19) is a global threat that raises worldwide concerns with escalating economic, social, and health challenges. On March 11, the World Health Organization (WHO) officially declared COVID-19 as a pandemic, pointing to the sustained risk of further global spread and urging countries and regions to join forces [1]. As of September 6 (the time of writing), there had been a total of 26,763,217 infections and 876,616 deaths globally, and the U.S. accounts for 23.0% of the global infections and 21.3% of the global deaths [2]. We are still witnessing widespread community transmission of the COIVD-19 all over the world. Unfortunately, to date, there is neither a vaccine nor a pharmacological agent found to reduce the transmission of severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), the virus that causes COVID-19 [3].

In response to the threat of COVID-19, social distancing measures are one of the primary tools to reduce the transmission of the SARS-CoV-2 virus. National and local governments have promoted stay-at-home orders while required non-essential business closures to reduce the risk of transmission by further enhancing social distancing measures [4]. Studies have found that such early governmental policies have been proved rather effective in China [5], Korea[6], and many European countries [7,8], as notable declines in transmission rates were observed following the implementation of strong mobility-reducing measures. In the U.S., many states, counties, and cities began issuing stay-at-home or similar mitigation measures that require residents to reduce movement and stay home as early as in March 2020, which leads to a considerable increase in home dwell time. The earliest stay-at-home order was implemented in the Bay Area, CA, on March 16, 2020, and soon after (3 days later), a state-wide stay-at-home order was issued in CA [9]. Gradually, an increasing number of states started to adopt this strategy. By March 24, more than 50% of the U.S. population was under a stay-at-home order, and this number soared from 50% to 95% by April 4 [9]. Despite the widely-adopted stay-at-home orders, there was mounting evidence of the disparate responses that potentially leaves vulnerable populations unequally exposed to the COVID-19 pandemic [10,11]. When facing the same mitigation measures, such disparate responses from residents are largely due to the different socioeconomic status, reflecting the long-standing problem of health inequity in the U.S., which usually exaggerates the consequences from disproportionate responses by inflicting long-term negative outcomes for the socially-disadvantaged groups [12]. Thus, identifying demographic and socioeconomic variables that potentially drive the disparity in the implementation of stay-at-home orders deserves much attention.

Many studies have investigated the disparity in response to the stay-at-home orders during the COVID-19 pandemic. The responses are usually quantified by home dwell time, travel distance, and POI (point of interest) visits, thanks to the availability of mobility datasets that facilitate the rapid monitoring of human mobility. Chiou and Tucker [13] investigated the U.S. tract-level correlation between income and self-isolation at home and found that high-income earners generally spend more time at home (their evidence points out that the access to high-speed Internet plays an important role). Barnett-Howell and Mobarak [14] found that people with less income tend to place greater value on their livelihood concerns than contracting COVID-19, consequently resulting in smaller epidemiological and economic benefits of social distancing measures in poorer regions. Jones [15] documented the urban-rural discrepancy in threat awareness, as 54% of urban residents in the states view the COVID-19 as a major threat, compared with 42% of those living in the suburbs and just 27% of rural residents in the same states. This urban-rural disparity in risk awareness presumably drives the disparity in their daily movement, therefore further translates to the disparity in COVID-19 exposure. Via the investigation of U.S. county-level mobility records, Lou et al. [16] found differential impacts of stay-at-home orders on economics groups, where the lower-income group is less likely to follow the order to stay at home, evidenced by their longer travel distances compared with the higher-income group. Similarly, Huang et al. [10] compared four popular mobility datasets and found that, regardless

of their unique characteristics, all selected mobility datasets suggest a statistically significant positive correlation between mobility reduction and income at the U.S. county scale. Despite the above efforts, the soundness of correlating disparity in response to demographic/socioeconomic variables is hampered by the coarse geographical units, as mitigation policies may vary in different countries, states, and even counties; therefore, the documented disparity in response may result from the discrepancy in mitigation policies, not from the varying demographic/socioeconomic indicators. Thus, the examination of fine-grained mobility records (e.g., at the census tract or block group level) are in great need. In addition, most existing studies utilize indices summarized during a specific period to quantify the mobility-related response, neglecting the dynamic perspectives revealed from time-series data. In comparison, time-series trend-based analytics may provide valuable insights in distinguishing different dynamic patterns of mobility records, thus warranting further investigation.

The objective of this study is to explore the capability of time-series clustering in categorizing fine-grained mobility records during the COVID-19 pandemic, and further investigate what demographic/socioeconomic variables differ among the categories with statistical significance. Taking advantage of the home dwell time at Census Block Group (CBG) level from the SafeGraph [17], and using the Atlanta-Sandy Springs-Roswell metropolitan statistical area (MSA) (hereafter referred to as Metro Atlanta) as a study case, this study investigates the potential driving factors that lead to the disparity in the time-series of home dwell time during the COVID-19 pandemic, providing fundamental knowledge that benefits policy-making for better mitigation measures of future pandemics. The contributions of this work are summarized as follows:

- We perform a trend-driven analysis by conducting Kmeans time-series clustering using fine-grained home dwell time records from SafeGraph.
- We assess the statistical significance of sixteen selected demographic/socioeconomic variables among categorized groups derived from the time-series clustering. Those variables cover economic status, races and ethnicities, age and household type, education, and transportation.
- We discuss the potential demographic/socioeconomic variables that lead to the disparity in home dwell time during the COVID-19 pandemic, how they reflect the long-standing health inequity in the U.S., and what can be suggested for better policy-making.

The remainder of the paper is organized as follows. Section 2 introduces the datasets used in this study. Section 3 presents the methodological approaches we applied. Section 4 describes the contexts of the study case (Metro Atlanta). Section 5 presents the results of time-series clustering, the results of the analysis of variance, and the discussion. Section 6 concludes our article.

2. Datasets

3.1. Home dwell time

The home dwell time records are derived from SafeGraph (https://www.safegraph.com/), a data company that aggregates anonymized location data from numerous applications in order to provide insights about physical places. SafeGraph aggregates data using a panel of GPS points from anonymous mobile devices and determines the home location as the common nighttime location of each mobile device over a six-week period to a Geohash-7 granularity (~153m × ~ 153m) [17]. To enhance privacy, SafeGraph excludes CBG information if fewer than five devices visited an establishment in a month from a given CBG. The data records used in this study are the median home dwell time in minutes for all devices with a certain CBG on a daily basis. For each device, the observed minutes at home across the day are summed, and the median value for all devices with a certain CBG is further calculated [17]. The raw SafeGraph dataset we used for the year 2020 spans from January 1, 2020, to August 31, 2020 (244 days) with daily home dwell records (in mins) for a total of 219,972 CBGs. Heat map of home dwell time for these CBGs are

presented in Figure 1. The impact of COVID-19 can be observed, as home dwell time notably increased after the declaration of National Emergency on March 13, 2020 [18] (Figure 1), despite the disparity in the increasing intensity. After the lifting of strict social distancing measures in early May, however, home dwell time starts to decrease and returns to the pre-pandemic level (Figure 1). The increased variation of home dwell time after the National Emergency declaration indicates that CBGs have different responses to the pandemic and the government order. Despite the large number of CBGs, not all CBGs contain sufficient records to derive stable time-series that can be used for clustering. The details of the preprocessing steps are presented in Section 3.1.

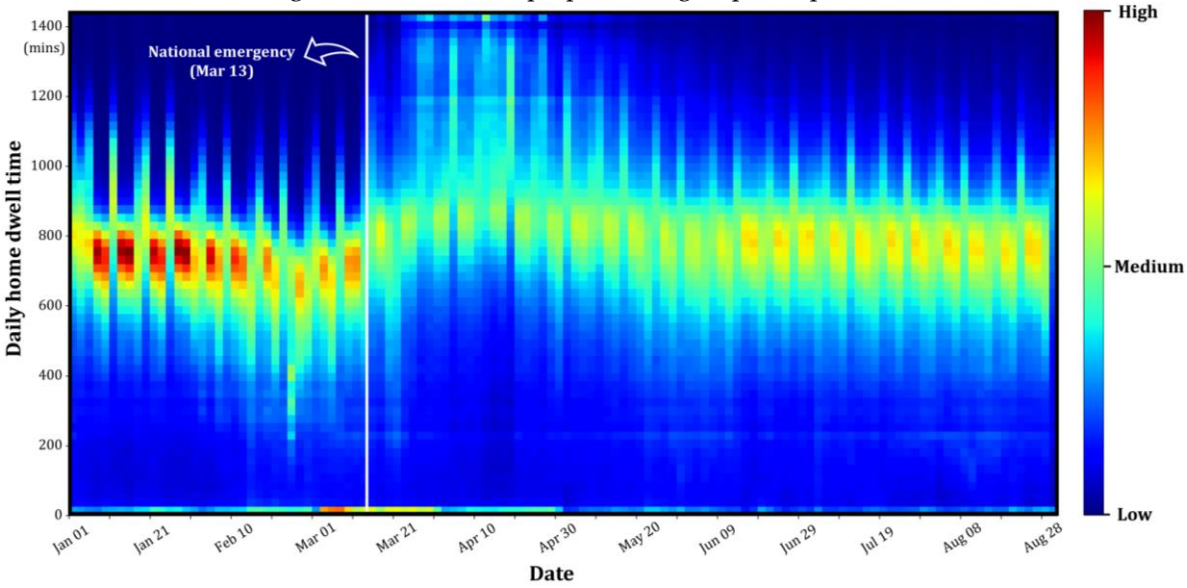

Figure 1. Heat map of home dwell time for 219,972 CBGs in the U.S. from January 1, 2020, to August 31, 2020. High/low concentrations are marked as red/blue.

3.2. Demographic/socioeconomic variables

Demographic and socioeconomic variables in this study are derived from the American Community Survey (ACS), collected by the U.S. Census Bureau. ACS is an ongoing nationwide survey that investigates a variety of aggregated information about U.S. residents at different geographic levels every year [19]. ACS randomly selects monthly samples based on housing unit addresses and publishes annual estimates datasets (i.e., 12-month samples). In addition to the 1-year datasets, ACS also releases 3-year estimates (i.e., 36-month samples) and 5-year estimates (i.e., 60-month samples). Compared to the 1-year and 3-year datasets, 5-year estimates cover the most areas, have the largest sample size, and contain the most reliable information [20]. In this study, we use the latest 5-year ACS data, i.e., the 2014-2018 ACS 5-year estimates, obtained from Social Explorer (https://www.socialexplorer.com/). We recode the variables from ACS data as five major categories: 1) Economic status; 2) Races and ethnicities; 3) Gender, age and household type; 4) Education; 5) Transportation. Previous empirical studies suggested that these variables could be associated with the pattern of daily travels and participation of out-of-home activities [21-24]. The detailed information of the variables within the five categories is presented in Table 1. In addition, CBG boundaries are derived from 2019 TIGER/Line Shapefiles by U.S. Census Bureau (https://www.census.gov/cgi-bin/geo/shapefiles/index.php).

Table 1. Notations and descriptions of the demographic and socioeconomic variable in the selected five major categories.

Variable notations	Descriptions
Economic status	
pct_low_income	percent of household income less than $15,000

pct_high_income	percent of household income greater than $150,000
median_hhinc	median household income
pct_unemployrate	unemployment rate
Races and ethnicities	
pct_white	percent of White
pct_black	percent of Black
pct_hispanic	percent of Hispanic
Gender, age and household type	
pct_female	percent of female
pct_elderly	percent of age 65 or older
pct_signleparent	percent of single-parent families among parenting families having children under 18
Education	
pct_low_edu	percent of education equal or less than highschool
pct_grad_edu	percent of education of master, professional, or doctoral degrees
Transportation	
pct_workhome	percent of work from home
pct_short_commute	percent of commuters with 10 min or shorter trips
pct_long_commute	percent of commuters with 40 min or longer trips
pct_0car	percent of 0 car households

3. Methods

3.1. Preprocessing

Several preprocessing steps are applied to ensure that CBGs within the study area contain sufficient and valid records to derive stable time-series that can be used for clustering. We first select CBGs that fall within the study area, i.e., Metro-Atlanta (more details of the Metro-Atlanta can be found in Section 4), which results in a total of 2,687 CBGs. As SafeGraph uses digital devices to measure home dwell time, the number of available devices in each CBG greatly determines the representativeness and the stability of the time-series. We plot the spatial distribution of median daily device count within the Metro Atlanta area and observe that CBGs dominated by non-residential zones tend to have less daily device count (Figure 2a), presumably due to the low number of home locations identified via SafeGraph's algorithm (see Section 3.1). We keep CBGs with more than 200 days (out of 244 days) of home dwell time records to ensure reliable time-series can be generated. To fill the missing data, we adopt the approach from Huang et al. [10], where missing data are filled via a simple linear interpolation by assuming that home dwell time changes linearly between two consecutive available records. Our preliminary investigation suggests that stable time-series of daily home dwell time can be achieved when daily device count reaches 100. Thus, we calculate the median of daily device count for each CBG during the 244-day period and select CBGs with the median equal or larger than 100. We also observe that some CBGs present abnormal home dwell patterns with consecutive 0 values for a certain period of time. To avoid the potential problems caused by these CBGs on the performance of the clustering algorithm, we remove CBGs with 0 values that span more than three consecutive days. A total of 1,483 CBGs remain after the aforementioned preprocessing steps, and their representativeness is presented in Figure 2b. The representativeness is defined as the ratio between the median daily device count and the population from the ACS 2014-2018 estimates. The representativeness for most CBGs ranges from 5% - 10% (Figure 2b), which is considerably higher than Twitter [25], a commonly used open-sourced platform to derive mobility-related statistics.

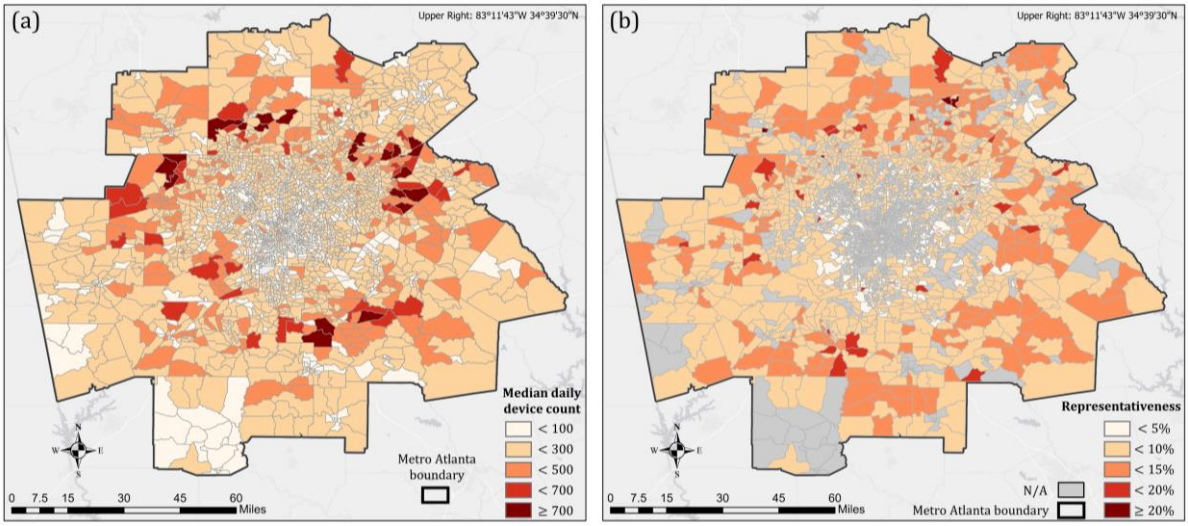

Figure 2. (a) The median daily device count from January 1 to August 31 at CBG level in Metro Atlanta; (b) Representativeness of each CBG in Metro Atlanta. The representativeness is defined as the ratio of the median daily device count to the population from the ACS 2014-2018 estimates. CBGs annotated with "N/A" fail to meet the requirements in the preprocessing steps and therefore are removed. CBG boundaries are derived from 2019 TIGER/Line Shapefiles.

3.2. Time-series clustering

Time-series clustering is the process of the partitioning a time-series dataset into a certain number of clusters, according to a certain similarity criterion. In this study, we aim to cluster the time-series of home dwell time in the CBGs within the study area. We adopt the design of K-means [26], an unsupervised partition-based clustering algorithm in which observations are categorized into the cluster with the nearest mean. The choice of similarity measurement in Kmeans is crucial to the detection of clusters [27]. Considering that the time-series of home dwell time for the majority of the CBGs present a similar shape but vary in intensity (Figure 1), we decide to calculate the Euclidean distance between two time-series.

Given a dataset on n time series $T = \{t_1, t_1, \dots, t_n\}$, we aim to partition T into a total of k clusters, i.e., $C = \{C_1, C_2, \dots C_k, \}$ by minimizing the objective function J, given as:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| t_i^j - C_j \right\| \tag{1}$$

where t_i^j denotes the time-series t_i in category j, and $\|\cdot\|$ denotes the similarity measurement that measures the distance between t_i^j and the cluster center of C_j. Let t_i and C_j each be a m-dimensional vector, where m equals the length of the time series (244 in this case). As Euclidean Distance is selected as similarity measurement in this study, $\left\| t_i^j - C_j \right\|$ can be rewritten as:

$$\left\| t_i^j - C_j \right\| = \sqrt{\sum_{k=1}^{m} (t_{ik}^j - C_{jk})^2} \tag{2}$$

Further, Kmeans utilizes an iterative procedure with the following steps to derive the final category for each time-series candidate:

1. Initialize k cluster centroids $C_1, C_2, \dots C_k$, arbitrarily.
2. Assign each time-series t_i to its correct cluster C_j, according to $argmin \left\| t_i^j - C_j \right\|$.
3. Update the centers C_j based on the new clusters.
4. Repeat steps 2 and 3 until convergence.

The Kmeans time-series clustering requires pre-specification of the total number of clusters (i.e., k), which inevitably introduces the subjective nature of deciding the constitution of reasonable clusters [28]. Through the investigation of the time-series dataset, we set $k = 3$, expecting to find three CBG clusters with different home dwell time patterns, following the stay-at-home order: 1) CBGs with a significant increase of home dwell time; 2) CBGs with a moderate increase of home dwell time; 3) CBGs with unnoticeable changes in home dwell time.

3.3. Analytical approaches

After the time-series clustering, three CBG clusters are therefore formed, each with a unique distribution pattern of daily home dwell time. Identifying the statistical difference in demographic/socioeconomic variables among these clusters facilitates a better understanding of what variables potentially lead to the disparity in home dwell time during the COVID-19 pandemic. Qualitatively, we label the CBG clusters, plot them spatially, and compare the spatial pattern of clusters with the spatial pattern of several major demographic/socioeconomic variables in the study area (see Figure 3 in Section 4). Quantitatively, we apply one-way ANOVA (Analysis of Variance) ($\alpha = 0.001$) [29] to assess the statistical significance of five major indicators (see Table 1) among categorized CBG groups derived from the time-series clustering. As ANOVA does not provide insights into particular differences between pairs of cluster means, we further conduct Tukey's test (α = 0.05, 0.01, 0.001) [30], a common and popular post-hoc analysis following ANOVA, to assess the statistical difference of demographic/socioeconomic variable between cluster pairs.

4. Profile of the study area

The study area defined in this study is referred to as Metro Atlanta, designated by the United States Office of Management and Budget (OMB) as the Atlanta–Sandy Springs–Alpharetta, Georgia (GA) Metropolitan Statistical Area (MSA). Metro Atlanta is the twelfth-largest MSA in the U.S. and the most populous metro area in GA [31]. The study area includes a total of 30 GA counties (listed in Table A) and has an estimated population of 5,975,424, according to the ACS 2014-2018 estimates. Metro Atlanta has grown rapidly since the 1940s. Despite its rapid growth, however, Metro Atlanta has shown widening disparities, including class and racial divisions, underlying the uneven growth and development, making it one of the metro regions with the most inequity [32-34]. It is the main reason why we chose this metro region to explore the disparity in responses to the COVID-19 pandemic. In the last few decades, the north metro area has absorbed most of the new growth, thanks to the northward shifting trend of the metro region's white population and the rapid office, commercial, and retail development [35]. After the increasingly unbalanced development in recent decades, Metro Atlanta started to present a distinct north-south spatial disparity in many demographic/socioeconomic variables (Figure 3). Compared to the south metro region, the north region is characterized by higher income (Figure 3a), higher white percentages (Figure 3b), higher education (Figure 3c), and higher percentages of work-from-home workers (Figure 3d).

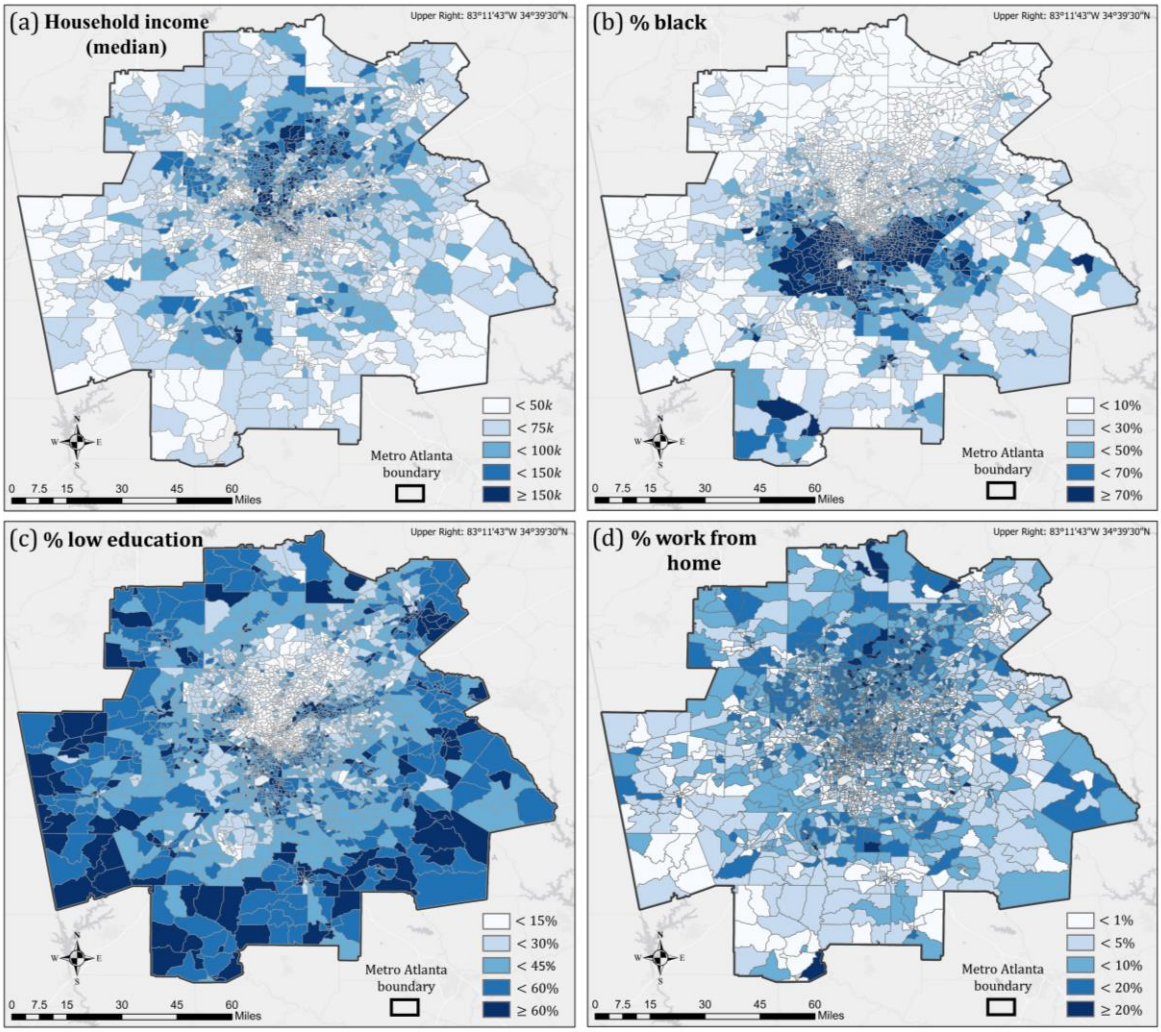

Figure 3. Profile of Metro Atlanta with four selected variables. (a) Median household income; (b) Percentage of black; (c) Percentage of low education (education equal or less than high school); (d) Percentage of workers who work from home. All statistics are derived from the ACS 2014-2018 estimates. CBG boundaries are derived from 2019 TIGER/Line Shapefiles.

In contrast to the substantial spatial heterogeneity of socioeconomic status, GA's governmental reactions to the COVID-19 pandemic are rather homogenous in space. On March 14, 2020, Governor Brian P. Kemp announced the public health state of emergency in GA. Twenty days later (April 3), the shelter-in-place order took effect for the entire state [36]. The strict social distancing measures lasted until late April when GA started to reopen gradually: resuming restaurant dine-in services (April 27), reopening bars and nightclubs with capacity limits (June 1), allowing the gatherings of 50 people (June 16), and reopening conventions and live performance (July 1) [37].

5. Results

5.1 Identified CBG clusters and their spatial distribution

Three CBG clusters are identified based on the time-series pattern in daily home dwell time via the Kmeans time-series clustering algorithm. CBGs in Cluster #1 are characterized by their unnoticeable changes in home dwell time throughout the entire time frame, suggesting that stay-at-home orders have a minimal effect on people living in these CBGs (Figure 4a). In comparison, CBGs in Cluster #2 (Figure 4b) and Cluster #3 (Figure 4c) responded to the strict measures

implemented in March and April strongly. CBGs in Cluster #2 experienced a moderate increase in home dwell time during the implementation of strict social distancing measures (Figure 4b). Compared to Cluster #2 where the daily home dwell time increased up to 1,200 mins, CBGs in Cluster #3 saw a more dramatic increase, as the home dwell time for most of the CBGs in Cluster #3 reached 1,400 mins (out of 1440 mins in a day) in March and April, suggesting that mitigation measures have greatly changed people's travel behavior in these CBGs (Figure 4c). Note that the three identified clusters are with different numbers of CBGs. Clusters #1, #2, and #3 have 157 CBGs, 778 CBGs, and 552 CBGs, respectively.

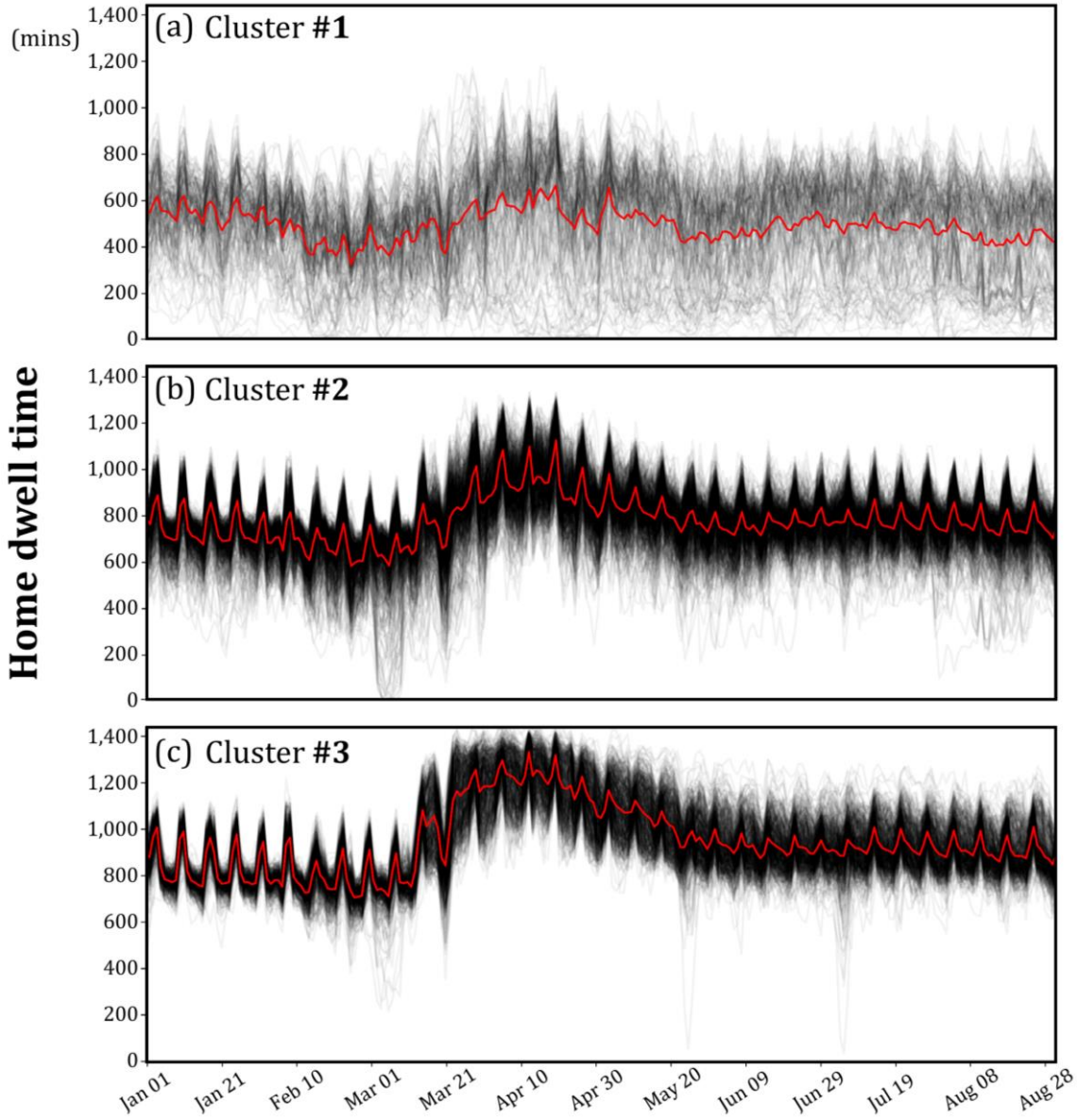

Figure 4. The time-series of the three identified CBG clusters. (a) Cluster #1: CBGs with unnoticeable changes in home dwell time (157 CBGs); (b) Cluster #2: CBGs with a moderate increase of home dwell time (778 CBGs); (c) Cluster #3: CBGs with a strong increase in home dwell time (552 CBGs).

Figure 5 shows the spatial distribution of the three CBG clusters, which presents a certain level of spatial autocorrelation, especially for Cluster #2 and Cluster #3. The Global Moran's I [38] for the distribution of the three identified clusters is 0.243, and it is significant at the significance level of 0.01. In general, the spatial distribution implies that demographic/socioeconomic variables potentially drive the disparity in home dwell time during the pandemic. The

distribution of CBGs in Cluster #3 suggests a high correlation of home dwell time and income, as the distribution patterns between CBGs in Cluster #3 and CBGs of high household income (see Figure 3a) are largely similar. North Metro Atlanta, where CBGs with high percentages of work-from-home workers and high educational levels are concentrated, exhibits a strong influence due to the stay-at-home orders, evidenced by the high concentration of CBGs in Cluster #3, a cluster with significantly increased home dwell time in March and April.

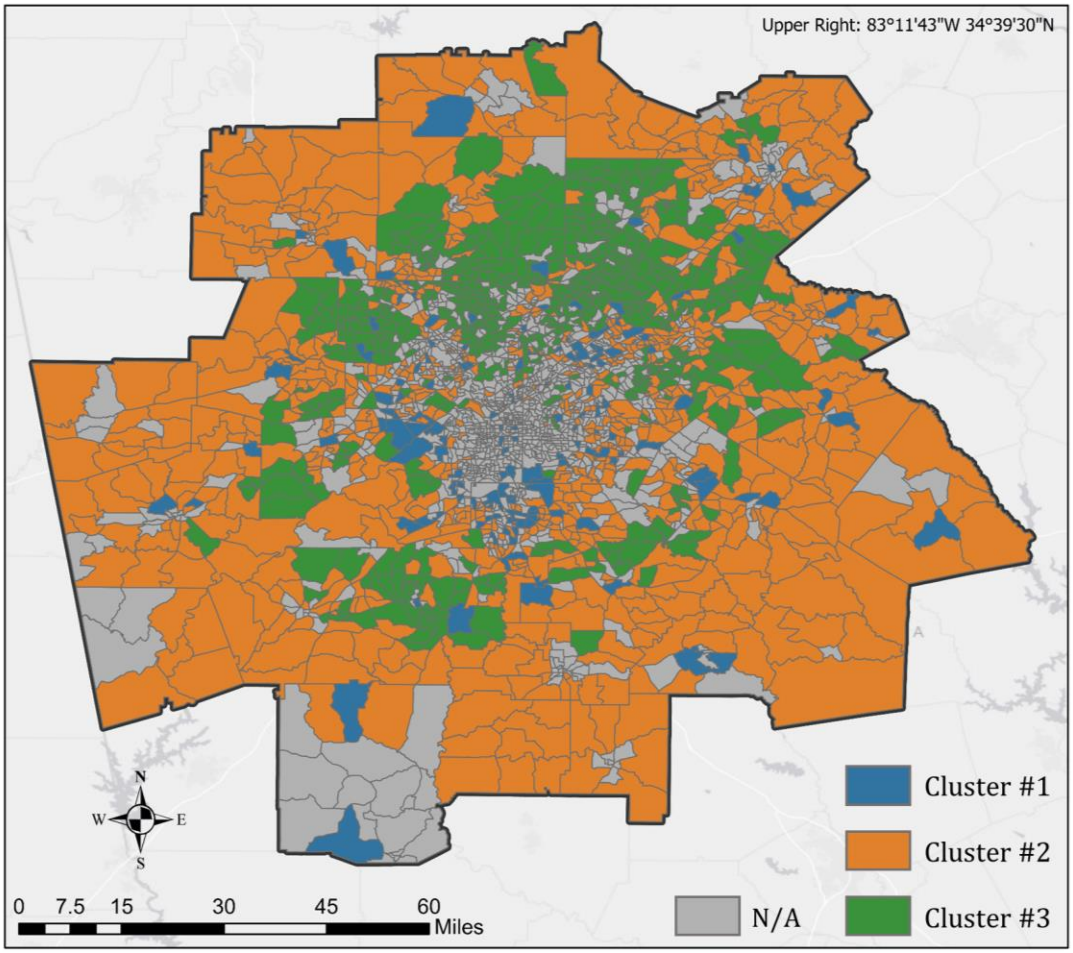

Figure 5. The spatial distribution of the three identified CBG clusters.

5.2 Demographic/socioeconomic variables in three identified clusters

The selected sixteen demographic/socioeconomic variables present unique distribution patterns in the three identified clusters (Figure 6). Compared with the other two clusters, Cluster #3 is characterized by a high median household income, a high percentage of high-earning groups, a low percentage of low-earning groups, and a low unemployment rate, suggesting that residents in rich CBGs respond to the stay-at-home order more aggressively by considerably reducing their out-of-home activities. It indicates that financial resources can, to a certain degree, influence the effectiveness of policies, as stated in other studies [16, 39].

In terms of racial composition, the three clusters are distinctly different. The mean Black percentages of CBGs in Cluster #1, #2, and #3 are respectively 49.5%, 31.3%, and 14.5%. CBGs in Cluster #1 (with unnoticeable home dwell time increase) present much higher Black percentages than Cluster #3 (with strong home dwell time increase), revealing that stay-at-home order is less effective for CBGs with higher Black percentages. This finding coincides with other recent studies that identified the racial disparities during the COVID-19 pandemic [40, 41]. As expected, Cluster #1 also presents a higher single-parent family percentage, given the fact that a high percentage of single-parent families is usually seen in Black communities [42]. In contrast, the three identified

clusters present similar Hispanic and female percentages, indicating their weaker role in distinguishing the patterns of home dwell time.

As for education, CBGs in Cluster #1 and #2 show similar distribution of the percentages of low education (42.1% and 43.1% as mean) while CBGs in Cluster #3 shows a considerably lower percentage (22.7% as mean). A reversed pattern can be found for high education, where Cluster #3 presents a notably higher percentage of high education compared to Cluster #1 and #2.

The percentages of short-commuters remain similar in all three clusters, while the percentages of long-commuters differ. The mean percentages of long-commuters in Cluster #1, #2, and #3 are 27.3%, 31.8%, and 35.4%, respectively. The result points out that a stronger increase in home dwell time is in tandem with a higher percentage of long-commuters.

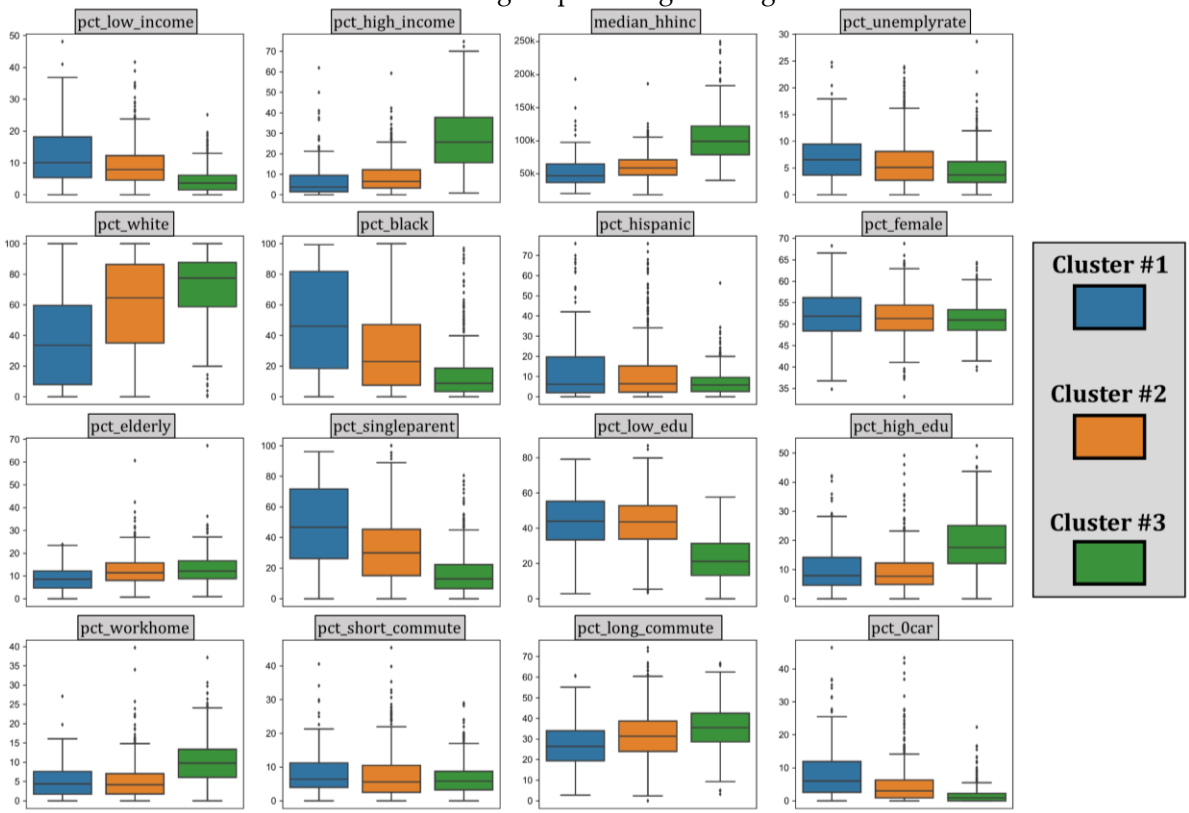

Figure 6. Selected demographic/socioeconomic variables in three identified clusters. The descriptions of these variables can be found in Table 1.

5.3 ANOVA and Tukey's test for clustered CBGs

We perform AVNOA to assess the statistical difference of demographic/socioeconomic variables among the three identified clusters and post-hoc Tukey's test to evaluate the statistical difference between a certain cluster pair. The results from ANOVA suggest that all selected variables, except for the percentage of females (pct_female) and the percentage of short-commuters (pct_short_commute), show a statistically significant difference ($\alpha = 0.001$) among the three clusters (Table 2). The results reveal that gender and the percentage of short-commuters are not significantly different ($\alpha = 0.001$) among the means of the three identified clusters, indicating that these two variables play a weaker role in explaining the disparity in patterns of home dwell time.

Table 2. ANOVA results.

Variable notations	F statistics	Sig. (*p*-value)
Economic status		
pct_low_income	151.71	<0.001*

pct_high_income	516.20	<0.001*
median_hhinc	507.33	<0.001*
pct_unemployrate	32.94	<0.001*
Races and ethnicities		
pct_white	108.15	<0.001*
pct_black	137.81	<0.001*
pct_hispanic	30.34	<0.001*
Gender, age, and household type		
pct_female	4.98	0.017
pct_elderly	22.40	<0.001*
pct_signleparent	177.58	<0.001*
Education		
pct_low_edu	366.83	<0.001*
pct_grad_edu	261.75	<0.001*
Transportation		
pct_workhome	193.02	<0.001*
pct_short_commute	4.56	0.022
pct_long_commute	36.19	<0.001*
pct_0car	131.14	<0.001*

*p<0.001

To provide deeper insights into the comparisons of selected variables between a specific pair of clusters, we further conduct post-hoc Tukey's test (Figure 7). For variables regarding economic status, Cluster #3 is statistically different (α = 0.001) from Cluster #1 and #2 in all four economic-related variables, i.e., pct_low_income, pct_high_income, median_hhinc, and pct_unemployrate. Cluster #1 and Cluster #2 present a weaker difference (α = 0.01) in median_hhinc and are not significantly different in pct_high_income. Results of racial and ethnic variables suggest that three clusters are statistically different from each other in pct_white, pct_black, and pct_hispanic, despite the weaker difference in pct_hispanic (α = 0.05) between Cluster #1 and #2. The difference in education (pct_low_edu and pct_high_edu) is not significant between Cluster #1 and Cluster #2 but is significant (α = 0.001) when comparing Cluster #3 to either Cluster #1 or #2. It suggests that CBGs in Cluster #3, a cluster with a strong increase in home dwell time, are characterized by their residents with high education, which is statistically different from the other two clusters. In addition, the three clusters are statistically different (α = 0.001) from each other in terms of long-commuters (pct_long_commute) and car ownership (pct_0car), suggesting that these two variables partially explain the disparity in home dwell time.

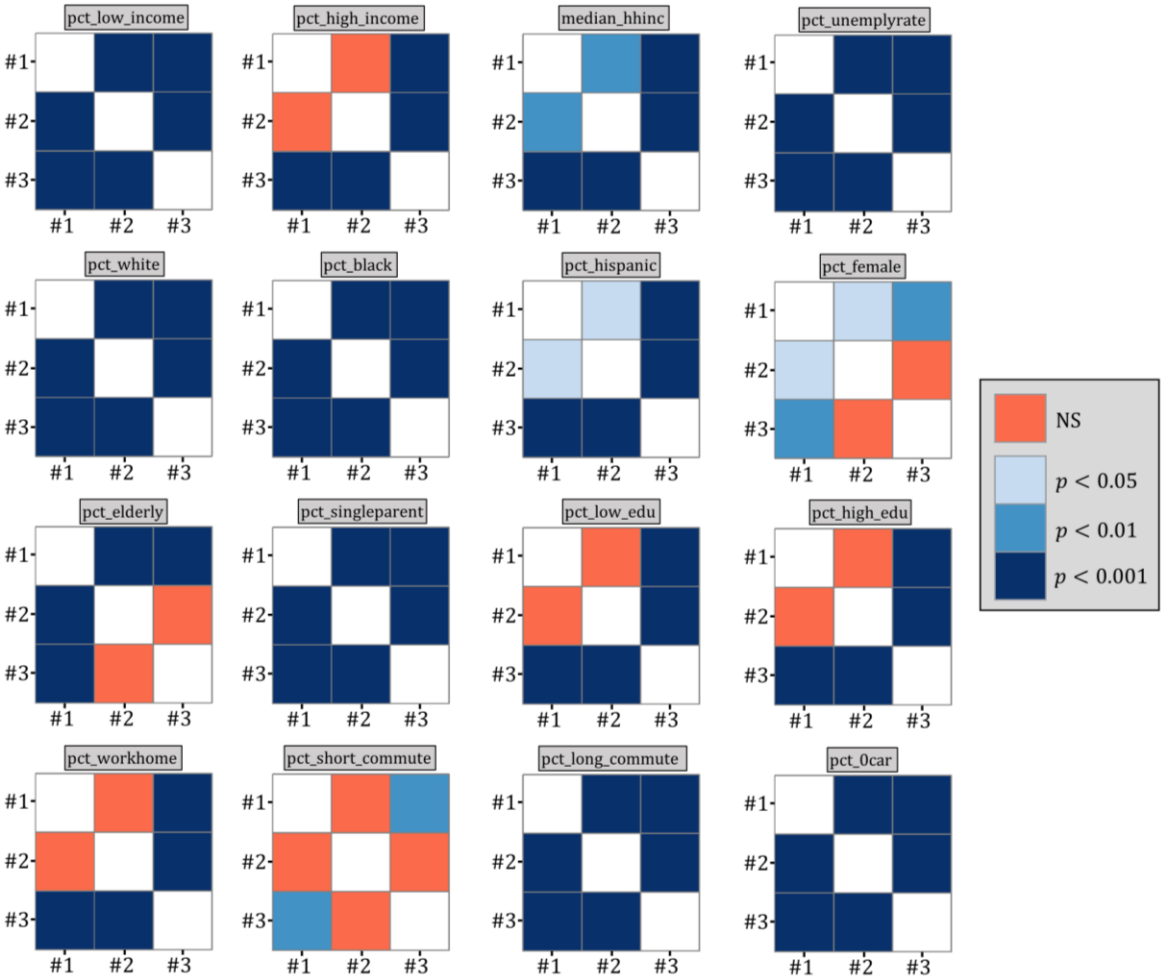

Figure 7. Post-hoc Tukey's test for selected variables between pairs of clusters. NS demotes "not significant" ($p \geq 0.05$).

6. Discussion

6.1 What do we learn?

This study applies a time-series clustering technique to categorize fine-grained mobility records (at CBG level) during the COVID-19 pandemic. Through the investigation of the demographic/socioeconomic variables in identified time-series clusters, we find that they are able to explain the disparity in home dwell time in response to the stay-at-home order, which potentially leads to disproportionate exposures to the risk from the COVID-19. This study also reveals that socially disadvantaged groups are less likely to follow the order to stay at home, pointing out the extensive gaps in the effectiveness of social distancing measures exist between socially disadvantaged groups and others. To make things worse, the existing socioeconomic status induced disparities are often exaggerated by the shortcomings of U.S. protection measures (e.g., health insurance, minimum incomes, unemployment benefits), potentially causing long-term negative outcomes for the socially disadvantaged populations [10]. In addition to the many pieces of epidemiological evidence that prove a strong relationship between social inequality and health outcomes [43, 44], this study offers evidence in the COVID-19 pandemic we are facing.

Specifically, we find that all selected variables, except for the percentage of females (pct_female) and the percentage of short-commuters (pct_short_commute), show a statistically significant difference ($\alpha = 0.001$) among the three identified clusters. CBGs in Cluster #3, a cluster with strong response in home dwell time, are characterized by high median household income, high Black percentage, high percentage of high-earning groups, low unemployment rate, high

education, low percentage of single parents, high car ownership, and high percentage of long-commuters. The statistically significant difference of demographic/socioeconomic variables in Cluster #3 collectively points out the privilege of the advantaged groups, usually the White and the affluent.

The weak response from the socially disadvantaged groups in home dwell time can be possibly explained by the fact that policies can sometimes unintentionally create discrimination among groups with different socioeconomic status [16], as people can react to policies based on the financial resources they have [45], which in return, influences the effectiveness of the policies. Our study reveals that the long-standing inequity issue in the U.S. stands in the way of the effective implementation of social distancing measures. Thus, policymakers need to carefully evaluate the inevitable trade-off among different groups and make sure the outcomes of their policies reflect not only the preferences of the advantaged but also the interests of the disadvantaged.

6.2 Limitations and future directions

It is important to mention several limitations of this study and provide guidelines for future directions. First, we acknowledge the subjectivity of predefining the number of clusters in the Kmeans clustering algorithm. In this study, we set the number of clusters as three (i.e., $k = 3$) via the investigation and interpretation of the home dwell time records from SafeGraph. We notice that, even after the preprocessing, some CBGs still present unstable temporal patterns due to the low and varying daily device count. Our interpretation of the data records reveals three distinct temporal patterns with a strong, moderate, and unnoticeable increase in home dwell time during March and April (hence, k is predefined as 3). To ensure the interpretability of clusters, the selection of the number of clusters in Kmeans via prior knowledge (priori) is common. However, we acknowledge that approaches like Elbow Curve [46] and Silhouette analysis [47] are largely adopted to facilitate the optimization of k without prior knowledge. When conducting a cross-city comparison or reproducing our approach in another region, we advise re-investigating the pattern of the time-series or adopting the aforementioned approaches to derive a reasonable setting of k.

Second, we construct and cluster the time-series of home dwell time using the data in the year 2020 (January 1 to Aug 31), without considering the changes in time-series compared to the previous year. It is reasonable to assume that deriving a cross-year change index facilitates the identification of CBGs that behave differently compared to the year 2019. However, we need to acknowledge the involvement of data records in the year 2019 inevitably introduces a certain level of uncertainty, as daily device count may vary substantially, leading to different representativeness of the same CBG between the two years. In addition, the Kmeans time-series clustering algorithm in this study takes the 8-month period as input. Further efforts can be directed towards the exploration of how CBGs behave differently at a certain time frame window, e.g., March and April, when strict social distancing measures were implemented.

Third, this study selects a total of sixteen variables from five major categories and explores the distribution of these variables in three identified clusters. Although previous studies have demonstrated the strong linkage between these variables and the participation of out-of-home activities, we can not rule out the possible contribution of other demographic/socioeconomic variables that are not included in this study. Future studies need to incorporate more variables to understand their roles in how social distancing guidelines are practiced. In addition, it is reasonable to assume that these variables drive the disparity in home dwell time, not independently but collectively. Therefore, statistical approaches like multinomial logit regression [48] can be used to further investigate the interactions among these variables towards time-series-based cluster generation.

Finally, it should be noted that the demographic structure, spatial pattern, and built environment vary substantially across areas, especially across densely populated urban fabrics

[49,50]. Thus, the influence of demographic/socioeconomic variables on the disparity in home dwell time following the stay-at-home order may not hold the same and tend to vary geographically. In addition, local governments had differing responses to the pandemic with varying strictness of the implemented social distancing measures, potentially leading to an unequal impact that disfavors disadvantaged groups. This study only explores the situation in Metro Atlanta, which can not be generalized to other regions without caution. Thus, it is necessary to conduct comparative studies that include multiple regions to better understand the contribution of demographic/socioeconomic variables to the impact of the COVID-19 pandemic on mobility-related behaviors.

7. Conclusion

This study categorizes the time-series of home dwell time records during the COVID-19 pandemic, and further explores what demographic/socioeconomic variables differ among the categories with statistical significance. Taking the Atlanta-Sandy Springs-Roswell metropolitan statistical area (Metro Atlanta) as a study case, we investigates the potential driving factors that lead to the disparity in the time-series of home dwell time, providing fundamental knowledge that benefits policy-making for better mitigation measures of future pandemics.

We find that demographic/socioeconomic variables can explain the disparity in home dwell time in response to the stay-at-home order, which potentially leads to disproportionate exposures to the risk from the COVID-19. The results further suggest that socially disadvantaged groups are less likely to follow the order to stay at home, pointing out the extensive gaps in the effectiveness of social distancing measures exist between socially disadvantaged groups and others. Specifically, we find that CBGs with strong response to the stay-at-home order are characterized by high median household income, high Black percentage, high percentage of high-earning groups, low unemployment rate, high education, low percentage of single parents, high car ownership, and high percentage of long-commuters, pointing out the privilege of the advantaged groups, usually the White and the affluent. In other words, populations with lower socioeconomic status may lack the freedom or flexibility to stay at home, leading to the exposure of more risks during the pandemic. Our study reveals that the long-standing inequity issue in the U.S. stands in the way of the effective implementation of social distancing measures. Thus, policymakers need to carefully evaluate the inevitable trade-off among different groups and make sure the outcomes of their policies reflect not only the preferences of the advantaged but also the interests of the disadvantaged.

Author Contributions: Conceptualization, Xiao Huang, Sicheng Wang, Hanxue Wei, and Baixu Cheng; Methodology, Xiao Huang, Junyu Lu, Sicheng Wang, Hanxui Wei, and Baixu Chen; Formal Analysis, Xiao Huang, Junyu Lu; Writing-Original Draft Preparation, Xiao Huang, Sicheng Wang, Hanxue Wei; Writing-Review & Editing, Xiao Huang, Junyu Lu, Zhenlong Li; Supervision, Xiao Huang, Zhenlong Li; Funding Acquisition, Xiao Huang.

Funding: This research was funded by the Vice Chancellor for Research & Innovation of the University of Arkansas.

Acknowledgments: The authors want to thank SafeGraph for providing the home dwell time dataset, which makes this research possible.

Conflicts of Interest: The authors declare no conflict of interest. The funders and data providers had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

Table A. U.S. counties that are included in Metro Atlanta.

GEOID[1]	County name
13171	Lamar County
13063	Clayton County
13089	DeKalb County
13227	Pickens County
13045	Carroll County
13297	Walton County
13013	Barrow County
13223	Paulding County
13199	Meriwether County
13113	Fayette County
13143	Haralson County
13015	Bartow County
13139	Hall County
13077	Coweta County
13159	Jasper County
13151	Henry County
13085	Dawson County
13097	Douglas County
13211	Morgan County
13135	Gwinnett County
13231	Pike County
13247	Rockdale County
13121	Fulton County
13149	Heard County
13067	Cobb County
13057	Cherokee County
13217	Newton County
13255	Spalding County
13035	Butts County
13117	Forsyth County

[1]GEOID is a 5-digit county code defined by the U.S. Census Bureau.

References

1. WHO Coronavirus Disease (COVID-19) - events as they happen. Available online: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen (assessed on 7 September 2020)
2. WHO Coronavirus Disease (COVID-19) – Weekly Epidemiological Update. Available online: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200907-weekly-epi-update-4.pdf?sfvrsn=f5f607ee_2 (assessed on 7 September 2020).
3. Le, T.T.; Andreadakis, Z.; Kumar, A., Roman; R.G., Tollefsen, S.; Saville, M.; Mayhew, S. The COVID-19 vaccine development landscape. *Nat. Rev. Drug. Discov.* **2020**, 19, 305-306.
4. Weill, J.A.; Stigler, M.; Deschenes, O.; Springborn, M.R. Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. *Proc. Natl. Acad. Sci.* **2020**, *117*, 19658-19660.
5. Kraemer, M.U.; Yang, C.H.; Gutierrez, B.; Wu, C.H.; Klein, B.; Pigott, D.M.; Du Plessis, L.; Faria, N.R.; Li, R.; Hanage, W.P.; Brownstein, J.S. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* **2020**, *368*, 493-497.
6. Shim, E.; Tariq, A.; Choi, W.; Lee, Y.; Chowell, G. Transmission potential and severity of COVID-19 in South Korea. *Int. J. Infect. Dis.* 2020, 93, 339-344.
7. Remuzzi, A.; Remuzzi, G. COVID-19 and Italy: what next? Lancet **2020**, 395, 1225-1228.
8. Stoecklin, S.B.; Rolland, P.; Silue, Y.; Mailles, A.; Campese, C.; Simondon, A.; Mechain, M.; Meurice, L.; Nguyen, M.; Bassi, C.; Yamani, E. First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures. *Eurosurveillance* **2020**, 25, 2000094.
9. Baek, C.; McCrory, P.B.; Messer, T.; Mui, P. Unemployment effects of stay-at-home orders: Evidence from high frequency claims data. *Institute for Research on labor and employment working paper* **2020**, 101-20.
10. Huang, X.; Li, Z.; Jiang, Y.; Ye, X.; Deng, C.; Zhang, J; Li, X. The characteristics of multi-source mobility datasets and how they reveal the luxury nature of social distancing in the U.S. during the COVID-19 pandemic. medRxiv **2020**, doi: 10.1101/2020.07.31.20143016.
11. Almagro, M.; Orane-Hutchinson, A. The Determinants of the Differential Exposure to COVID-19 in New York City and Their Evolution Over Time. *Covid Economics: Vetted and Real-Time Papers* **2020**.
12. Bonaccorsi, G.; Pierri, F.; Cinelli, M.; Flori, A.; Galeazzi, A.; Porcelli, F.; Schmidt, A.L.; Valensise, C.M.; Scala, A.; Quattrociocchi, W.; Pammolli, F. Economic and social consequences of human mobility restrictions under COVID-19. *Proc. Natl. Acad. Sci.* **2020**, 117, 15530-15535.
13. Chiou, L.; Tucker, C., 2020. Social distancing, internet access and inequality (No. w26982). National Bureau Econ. Res. **2020**, 10.3386/w26982.
14. Barnett-Howell, Z.; Mobarak, A.M. The Benefits and Costs of Social Distancing in Rich and Poor Countries. *arXiv preprint* **2020**, arXiv:2004.04867.
15. Urban residents in states hit hard by COVID-19 most likely to see it as a threat to daily life. Available online: https://www.pewresearch.org/fact-tank/2020/03/20/urban-residents-in-states-hit-hard-by-covid-19-most-likely-to-see-it-as-a-threat-to-daily-life/ (accessed on 13 September 2020)
16. Lou, J.; Shen, X.; Niemeier, D. Are stay-at-home orders more difficult to follow for low-income groups? Working Paper 2020.
17. SafeGraph-Social Distancing Metrics. Available online: https://docs.safegraph.com/docs/social-distancing-metrics (accessed on 14 September 2020).
18. Proclamation on Declaring a National Emergency Concerning the Novel Coronavirus Disease (COVID-19) Outbreak. Available online: https://www.whitehouse.gov/presidential-actions/proclamation-declaring-national-emergency-concerning-novel-coronavirus-disease-covid-19-outbreak/.
19. American Community Survey Information Guide. Available online: https://www.census.gov/content/dam/Census/programs-surveys/acs/about/ACS_Information_Guide.pdf (assessed on 15 September 2020).
20. When to Use 1-year, 3-year, or 5-year Estimates. Available online: https://www.census.gov/programs-surveys/acs/guidance/estimates.html (assessed on 15 September 2020).

21. Morency, C.; Paez, A.; Roorda, M. J.; Mercado, R.; Farber, S., 2011. Distance traveled in three canadian cities: Spatial analysis from the perspective of vulnerable population segments. *J. Transp. Geogr.* **2011**, 19, 39-50.
22. Farber, S.; Páez, A.; Mercado, R. G.; Roorda, M.; Morency, C. A time-use investigation of shopping participation in three Canadian cities: Is there evidence of social exclusion? *Transportation*, **2011**, 38, 17-44.
23. Farber, S.; Páez, A. My car, my friends, and me: A preliminary analysis of automobility and social activity participation. *J. Transp. Geogr.* **2009**, 17, 216-225.
24. Páez, A.; Gertes Mercado, R.; Farber, S.; Morency, C.; Roorda, M. Relative accessibility deprivation indicators for urban settings: Definitions and application to food deserts in Montreal. *Urban Stud.* **2010**, 47, 1415-1438.
25. Huang, X.; Li, Z.; Jiang, Y.; Li, X.; Porter, D. Twitter, human mobility, and COVID-19. *arXiv preprint* **2020**. arXiv:2007.01100.
26. Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; Wu, A. Y. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, 24, 881-892.
27. Liao, T. W. Clustering of time series data—a survey. *Pattern Recognit*, **2005**, 38, 1857-1874.
28. Pham, D. T.; Dimov, S. S.; Nguyen, C. D. Selection of K in K-means clustering. *Proc. Inst. Mech. Eng. C.* **2005**, 219, 103-119.
29. St, L.; Wold, S. Analysis of variance (ANOVA). *Chemom. Intell. Lab. Syst.* **1989**, 6, 259-272.
30. Abdi, H.; Williams, L. J. Tukey's honestly significant difference (HSD) test. In Encyclopedia of research design; Salkind, N. J. SAGE Publication, Inc.:Thousand Oaks, CA, 2010; Volume 3, pp. 1-5.
31. Metropolitan and Micropolitan Statistical Areas Population Totals and Components of Change: 2010-2019. Available online: https://www.census.gov/data/tables/time-series/demo/popest/2010s-total-metro-and-micro-statistical-areas.html (assessed on 20 September 2020).
32. Bobo, L.; Johnson, J.; Oliver, M.; Farley, R.; Bluestone, B.; Browne, I.; ... Massagli, M. Multi-City Study of Urban Inequality, 1992-1994 [Atlanta, Boston, Detroit, and Los Angeles]. *Inter-University Consortium for Political and Social Research* **2020**, Ann Arbor, MI.
33. Wyczalkowski, C. K.; Welch, T.; Pasha, O. Inequities of Transit Access: The Case of Atlanta, GA. *Journal of Comparative Urban Law and Policy*, **2020**, 4, 654-681.
34. Bullard, R. D.; Johnson, G. S.; Torres, A. O. Sprawl Atlanta: social equity dimensions of uneven growth and development. **1999**. Atlanta, GA: Clark Atlanta University, The Environmental Justice Resource Center.
35. Keating, L. Atlanta: Race, Class, and Urban Expansion. Philadelphia: Temple University Press, **2001**; pp. 7-40.
36. Press Releases, Governor Brian P. Kemp – Office of the Governor. Available online: https://gov.georgia.gov/press-releases?field_press_release_type_target_id=All&page=17 (assessed on 20 September 2020).
37. Where states reopened and cases spiked after the U.S. shutdown, The Washington Post. Available online: https://www.washingtonpost.com/graphics/2020/national/states-reopening-coronavirus-map/ (assessed on 20 Septermber 2020).
38. Ord, J. K.; Getis, A. Local spatial autocorrelation statistics: distributional issues and an application. *Geogr. Anal.* **1995**, 27, 286-306.
39. Karaye, I. M.; Horney, J. A. The impact of social vulnerability on COVID-19 in the US: an analysis of spatially varying relationships. *Am. J. Prev. Med.* **2020**, 59, 317-325.
40. Holtgrave, D. R.; Barranco, M. A.; Tesoriero, J. M., Blog; D. S.; Rosenberg, E. S. Assessing racial and ethnic disparities using a COVID-19 outcomes continuum for New York State. *Ann. Epidemiol.* **2020**, 48, 9-14.
41. Laurencin, C. T.; McClinton, A. The COVID-19 pandemic: a call to action to identify and address racial and ethnic disparities. *J. Racial Ethn Health Disparities* **2020**, 1-5.
42. Bianchi, S. M. The changing demographic and socioeconomic characteristics of single parent families. *J. Marriage Fam.* **1994**, 20, 71-97.

43. Nguyen, V. K.; Peschard, K. Anthropology, inequality, and disease: a review. *Annu. Rev. Anthropol.* **2003**, 32, 447-474.
44. Muennig, P.; Franks, P.; Jia, H.; Lubetkin, E.; Gold, M. R. The income-associated burden of disease in the United States. Soc. Sci. Med. 2005, 61, 2018-2026.
45. Mechanic, D. Disadvantage, inequality, and social policy. *Health Aff.* **2002**, 21, 48-59.
46. Kodinariya, T. M.; Makwana, P. R. Review on determining number of Cluster in K-Means Clustering. *Int. J. Adv. Res. Comput. Sci. Manag. Stud.* **2013**, 1, 90-95.
47. Lletı, R.; Ortiz, M. C.; Sarabia, L. A.; Sánchez, M. S. Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Anal. Chim. Acta.* **2004**, 515, 87-100.
48. Böhning, D. Multinomial logistic regression algorithm. *Ann. Inst. Stat. Math.* **1992**, 44, 197-200.
49. Nasri, A.; Zhang, L. Impact of metropolitan-level built environment on travel behavior. *Transp. Res. Rec.* **2012**, 2323, 75-79.
50. Zhang, L.; Hong, J.; Nasri, A.; Shen, Q. How built environment affects travel behavior: A comparative analysis of the connections between land use and vehicle miles traveled in US cities. *J. Transp. Land Use* **2012**, 5, 40-52.