Original Paper

Identifying the Socioeconomic, Demographic, and Political Determinants of Social Mobility and Their Effects on COVID-19 Cases and Deaths: Evidence From US Counties

Niloofar Jalali[1], MSc, PhD; N Ken Tran[2], MSc, PhD; Anindya Sen[3], PhD; Plinio Pelegrini Morita[2,4,5,6], MSc, PEng, PhD

[1]School of Public Health and Health Systems, Faculty of Applied Health Sciences, University of Waterloo, Waterloo, ON, Canada

[2]School of Public Health and Health Systems, University of Waterloo, Waterloo, ON, Canada

[3]Department of Economics, University of Waterloo, Waterloo, ON, Canada

[4]Department of Systems Design Engineering, University of Waterloo, Waterloo, ON, Canada

[5]JW Graham Information Technology Emerging Leader Chair, Applied Health Informatics, University of Waterloo, Waterloo, ON, Canada

[6]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada

Corresponding Author:
Plinio Pelegrini Morita, MSc, PEng, PhD
School of Public Health and Health Systems
University of Waterloo
200 University Avenue West
Waterloo, ON, N2L 3G1
Canada
Phone: 1 5198884567 ext 31372
Email: plinio.morita@uwaterloo.ca

Abstract

Background: The spread of COVID-19 at the local level is significantly impacted by population mobility. The U.S. has had extremely high per capita COVID-19 case and death rates. Efficient nonpharmaceutical interventions to control the spread of COVID-19 depend on our understanding of the determinants of public mobility.

Objective: This study used publicly available Google data and machine learning to investigate population mobility across a sample of US counties. Statistical analysis was used to examine the socioeconomic, demographic, and political determinants of mobility and the corresponding patterns of per capita COVID-19 case and death rates.

Methods: Daily Google population mobility data for 1085 US counties from March 1 to December 31, 2020, were clustered based on differences in mobility patterns using K-means clustering methods. Social mobility indicators (retail, grocery and pharmacy, workplace, and residence) were compared across clusters. Statistical differences in socioeconomic, demographic, and political variables between clusters were explored to identify determinants of mobility. Clusters were matched with daily per capita COVID-19 cases and deaths.

Results: Our results grouped US counties into 4 Google mobility clusters. Clusters with more population mobility had a higher percentage of the population aged 65 years and over, a greater population share of Whites with less than high school and college education, a larger percentage of the population with less than a college education, a lower percentage of the population using public transit to work, and a smaller share of voters who voted for Clinton during the 2016 presidential election. Furthermore, clusters with greater population mobility experienced a sharp increase in per capita COVID-19 case and death rates from November to December 2020.

Conclusions: Republican-leaning counties that are characterized by certain demographic characteristics had higher increases in social mobility and ultimately experienced a more significant incidence of COVID-19 during the latter part of 2020.

(JMIR Infodemiology 2022;2(1):e31813) doi: 10.2196/31813

KEYWORDS

COVID-19; cases; deaths; mobility; Google mobility data; clustering

Introduction

In March 2020, COVID-19 was acknowledged by the World Health Organization (WHO) to be a global pandemic [1]. Since then, governments worldwide have implemented a series of lockdown measures intended to reduce the spread of the disease. The efficacy of these measures, in the absence of a vaccine or effective therapy, has varied across countries. Initial evidence on lockdown measures implemented in China suggested that reducing interpersonal physical contact or reducing the movement of the population is an effective means to control the spread of the virus [2]. These findings spurred national and subnational policies restricting population mobility, including social distancing (physical distancing between people who are not from the same household) [3] and stay-at-home (SAH) or shelter-in-place (SIP) orders, which required people to stay at home except for essential activities [4,5].

In addition to the direct impacts of such policies, evaluating the effects of demographic and socioeconomic factors on population mobility is also important as there were non-pandemic-related events that significantly impacted public movements in the U.S. after the first wave of the pandemic. Specifically, the summer of 2020 witnessed many demonstrations and public rallies in the U.S. in response to a series of events, including the death of George Floyd. Social distancing receded into the background despite rising caseloads and deaths due to COVID-19. The initial decline in public movement that occurred during the early months of the pandemic was succeeded by rapid increases in social mobility through much of the U.S. [6]. Increases in social mobility also occurred as many jurisdictions modified their SAH orders, allowed more businesses to reopen, and relaxed rules on social distancing [7]. This rise in mobility has been linked to higher COVID-19 cases in these regions [8]. Public mobility may have also increased during fall 2020 because of public rallies and social gatherings associated with the US presidential election.

A growing amount of research has used mobility data from social media platforms (Google, Twitter, and Facebook) and mobile phone providers to understand changes in mobility during the pandemic [9,10], the relationship between population mobility and the spread of COVID-19 cases [8-18], and the effects of nonpharmaceutical interventions (NPIs) on mobility [5,19,20]. The consensus from these studies is that increased mobility is associated with higher COVID-19 case counts. Badr et al [15] used cell phone data from 25 counties provided by Teralytics and found that reduced mobility patterns are associated with reduced COVID-19 incidence rates. Using mobile phone data from Safegraph, Gao et al [20] similarly found that lower mobility (more time at home) is associated with a reduced spread of COVID-19 across states. Glaeser et al [19] also used Safegraph data and found reduced mobility to be correlated with lower cases for some US cities. Using Google data from different jurisdictions, other studies found a positive correlation between mobility and COVID-19 case counts [11,12,14,17]. These studies are, however, limited; they investigated social mobility across a small number of US counties during the early days of the pandemic. As such, they were unable to capture socioeconomic, demographic, and political determinants of mobility [21-25].

We evaluated the determinants and consequences of population movements in 1089 US counties from the start of the pandemic to December 2020. This study contributes to the literature by using clustering analysis and other tools to evaluate the impacts of different socioeconomic and demographic characteristics on social mobility in a sample of US counties. We also investigated the effects of such mobility decisions on daily per capita COVID-19 cases and deaths. Social mobility was measured through the use of Google mobility indicators at retail and recreational venues, grocery and pharmacy stores, workplaces, and residences. Robust statistical findings based on such analysis would inform policymakers in crafting efficient and effective NPIs that could curb the spread of COVID-19.

Our results demonstrate that clusters with higher mobility at retail outlets, grocery and pharmacy stores, and workplaces and a lower duration of stay at residences also have a higher percentage of population aged 65 years and over, a larger population share of Whites with less than high school and college education, a higher percentage of the population with less than a college education, a lower percentage of the population using public transit to work, and a smaller share of voters who voted for Clinton during the 2016 presidential election relative to other clusters. The clusters with higher mobility also experienced pronounced increases in per capita COVID-19 daily case and death rates from November to December 2020. These findings are consistent with other studies that suggest that Trump-leaning counties experienced increases in social mobility and less stringent policies after the first wave of the pandemic, which was succeeded by higher levels of disease severity during the latter months of 2020.

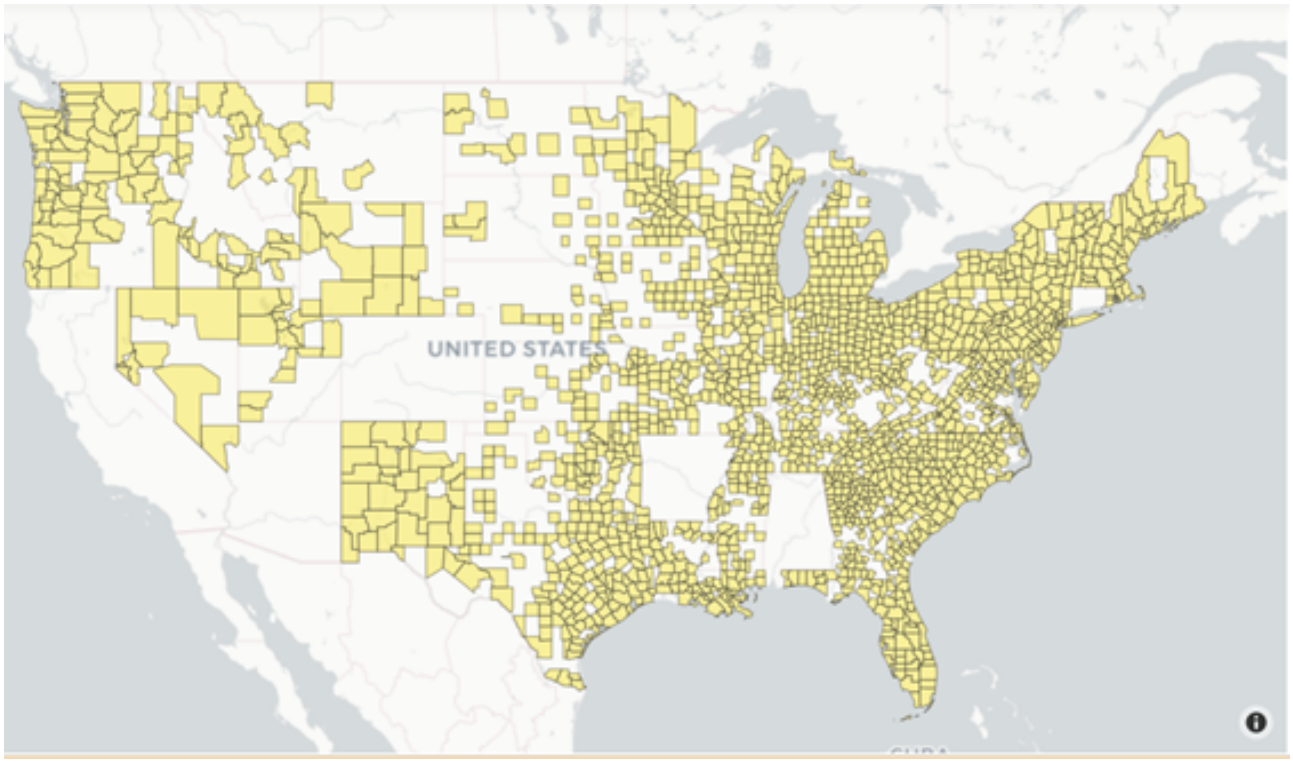
Methods

Data

COVID-19 Incidence Data

The daily numbers of confirmed cases and deaths due to COVID-19 at the county level were downloaded from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) [26]. For the 1089 counties in our sample, the mean (SD) of confirmed cases and deaths (both per 100,000 of population) were 1541.27 (1905.59) and 33.72 (44.78), respectively. Figure 1 reveals the distribution of counties in our sample. There is a significant concentration of counties in the East, Northeast, and certain southern states. There are fewer counties from the midwestern and southwestern parts of the United States. This is because Google mobility data (discussed later) are less available for counties with lower population density. This is a limitation of our analysis.

Figure 1. Geographic distribution of counties.



Population Mobility

Data on population mobility were obtained from Google's COVID-19 *Community Mobility Reports*. Google creates social mobility data from users who have turned on the Location History setting of Google accounts on their phones and have agreed to share this information. Google mobility indicators are with respect to population-level daily visits to grocery and pharmacy stores, which include grocery markets, food warehouses, farmers' markets, specialty food shops, drug stores, and pharmacies; parks, which consist of local parks, national parks, public beaches, marinas, dog parks, plazas, and public gardens; transit stations, comprising subway, bus, and train stations; retail stores and recreation outlets consisting of places such as restaurants, cafes, shopping centers, theme parks, museums, libraries, and movie theaters; and workplaces. The Google mobility data also provide an index on the duration of stay at residences. Google mobility indicators for transit hubs and parks were omitted because of large numbers of missing values for the counties included in this study.

A prepandemic baseline mobility value was determined using the median mobility for each day of the week from January 3 to February 6, 2020 [27]. Subsequent mobility values were normalized to baseline. Counties with missing values less than or equal to 10% for each indicator were selected for the study. Missing values were replaced by the average from 3 prior days. The availability of Google data determined which counties we used in our analysis. The final data set contained observations for 1089 counties, which is roughly 35% of the total number of counties (N=3142) in the United States. Daily values were available for the first and second waves of the pandemic from March 11 to December 31, 2020.

With the exception of the residential index, daily values for each index were calculated relative to baseline, which was defined as the median for the corresponding day of the week, during the 5-week period from January 3 to February 6, 2020. Hence, each daily value is the percentage change in the social mobility category relative to its baseline, which shows how the number of visits to different destinations in a day have changed in percentage terms since the onset of the pandemic. The Google residential index represents the duration of stay at an individual's residence relative to the 5-week baseline. The values in this index are the percentage differences in time spent at home relative to the baseline period.

County-Level Socioeconomic, Demographic, and Political Data

The 2016 census data were collected by the Massachusetts Institute of Technology (MIT) Election Data and Science Lab [18]. These data were supplemented by county variables collected by other studies [23,25]. To validate that our samples were representative of all US counties, we compiled summary statistics of socioeconomic and demographic variables between our sample and all counties (Table 1). In summary, there did not seem to be significant differences in most variables between all counties and our sample. The exception is population, where our sample mean was more than 2.5 times that of the mean for all counties. In a similar vein, although all counties have 58% of the population in rural areas, the corresponding statistic for our sample was only approximately 31%. These discrepancies can be explained by the fact that Google's social mobility indicators are only available for counties with larger populations that are more densely populated. This is consistent with the visualization of counties in our sample from Figure 1.

Table 1. Sample statistics of census variables for all counties and our sample based on daily values.

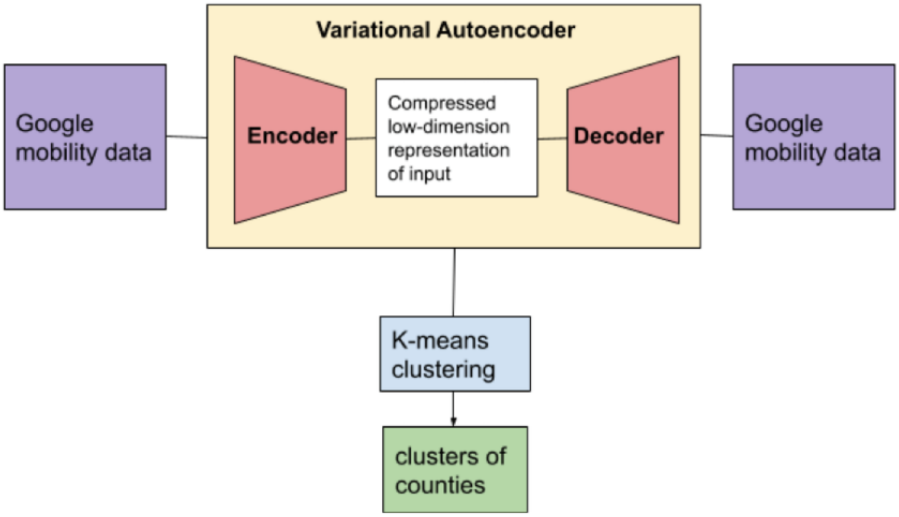
Variable	All counties			Our sample		
	Mean (SD)	Minimum	Maximum	Mean (SD)	Minimum	Maximum
Politics						
Population voting for Trump in 2016 (%)	28.13 (8.44)	1.93	76.32	24.28 (7.22)	2.63	66.42
Population voting for Clinton in 2016 (%)	14.07 (7.41)	0	49.02	17.18 (7.33)	2.73	42.86
Registered voters as population (%)	74.86 (5.31)	43.14	95.08	73.49 (5.14)	47.33	90.63
Demographics						
Whites (%)	77.36 (19.74)	0.76	100.00	73.57 (18.63)	2.78	97.34
African Americans (%)	8.96 (14.5)	0	86.19	9.96 (12.21)	0.09	76.55
Hispanics (%)	8.99 (13.66)	0	98.96	11.03 (13.41)	0.68	95.48
Foreign born (%)	4.62 (5.63)	0	52.23	7.12 (6.81)	0.40	52.23
Females (%)	49.98 (2.33)	21.51	58.50	50.62 (1.30)	38.76	56.03
Population aged 29 years and under (%)	37.34 (5.44)	11.84	70.98	39.24 (4.98)	13.64	61.69
Population aged 65 years and older (%)	17.63 (4.44)	3.86	53.11	15.57 (3.93)	6.95	53.11
Less than high school education (%)	14.23 (6.54)	1.28	51.48	12.44 (5.26)	2.08	41.34
Less than college education (%)	79.22 (9.14)	19.79	97.02	73.98 (10.11)	26.34	90.86
Whites with less than high school education (%)	11.04 (5.33)	0	41.76	9.11 (3.92)	0.97	25.57
Whites with less than college education (%)	77.00 (10.36)	9.19	95.92	71.28 (11.58)	15.30	89.96
Socioeconomics						
Median household income (US $)	47,817.60 (12482.4)	18,972.00	125,672.00	53,798.50 (13905.9)	28,452.00	125,672.00
Rural population (%)	58.48 (31.45)	0	100.00	31.733 (22.08)	0	100.00
Population density (number of people per square mile)	582.71 (3761.83)	0.26	179,922.30	1397.32 (6127.90)	6.22	179,922.30
Hospitals per 100,000 of population (number of hospitals per 100,000 of population)	0.61 (0.94)	0	10.56	0.25 (0.166)	0	1.61
Poverty rate (%)	15.16 (6.07)	2.60	48.40	13.35 (4.87)	2.60	37.30
Population without health insurance (%)	0.09 (0.05)	0.01	1.62	0.09 (0.06)	0.02	1.62
Share of population using public transit for commuting to work (%)	0 (0.01)	0	0.26	0.01 (0.02)	0	0.26

Clustering

Figure 2 summarizes our methodology for identifying different clusters of counties using Google mobility indicators. Clustering is an unsupervised learning technique that partitions a data set into groups or clusters based on similarity measures. This study leveraged partitioning-based algorithms, which divided the data set into partitions, where each partition was a cluster. For each county included in this study, data were clustered based on a combination of the daily values of the 4 Google mobility indicators. To identify the different clusters of counties, we performed 2 steps [28]:

1. Compressing the multidimensional time series data to extract the latent variables using deep neural networks
2. Using K-means clustering to identify the different clusters of counties based on latent variables' representations

Figure 2. Methodology for identifying different clusters of counties using a variational autoencoder.



To compress the multidimensional time series, we implemented the variational autoencoder (VAE) architecture based on long short-term memory (LSTM) [29-31]. The principal concept of this generative approach is to project high-dimensional data into latent variables. Our model comprised 4 blocks [32]:

1. Encoder: Defined by the LSTM layers, the multidimensional time series input (x) are fed into the LSTM.
2. Encoder to latent layer: Defined by a linear layer, which identifies the mean and SDs of the last hidden layer of the encoder. During the training process, the multigaussian distributions are defined and reparametrized iteratively by the mean and SDs derived from latent vectors.
3. Latent layer to decoder layer: The latent variables (z) are sampled from the distribution and pass through a linear layer to identify the decoder input.
4. Decoder: Defined by the LSTM layers, which uses latent variables (z) to reconstruct the original data [33].

Identifying the true posterior distribution is intractable [33]. Therefore, to construct the original data, the probabilistic encoder model was approximated by normal distribution p(z|x)N(0,1) and used as a probability decoder [30,33]. Hence, the reconstruction of input was defined by sampling from the distribution of latent variables (z).

To evaluate the performance of the model, the loss function was defined as follows:

- The divergence from the approximated distribution and the true distribution

$$D_{KL}[q(z|x) \,||\, \hat{q}(z|x)] = E[\log(q(z|x)) - \log\frac{p(X|z)p(z)}{P(x)}] \quad (1)$$

$$D_{KL}[N(\mu_x, \sigma_x^2)|| \, N(0,1)] = \frac{1}{2}\sum_k(\exp(\sigma_x^2) + \mu_x^2 - 1 - \sigma_x^2) \quad (2)$$

- The mean squared error loss calculated the difference between original and reconstructed input data

$$[\text{mse}(x - \hat{x})^2]$$

- The total loss is defined as sum of 2 losses:

$$\text{Loss} = \text{mse}(x - \hat{x})^2 + \frac{1}{2}\sum_k(\exp(\sigma_x^2) + \mu_x^2 - 1 - \sigma_x^2) \quad (3)$$

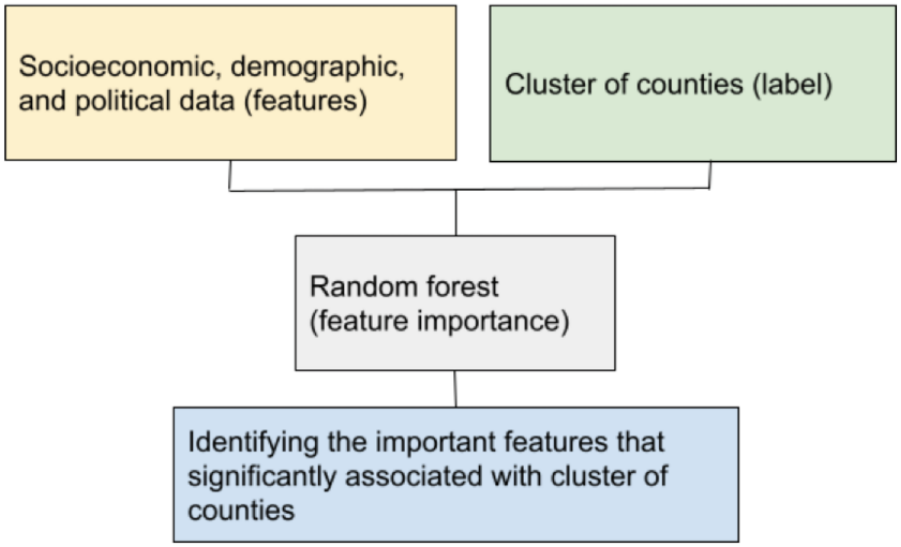
The model was trained in Python 3.6 using the Keras library [34] with the Adam optimizer. The batch size and number of the epochs were set to 10 and 100, respectively. The number of nodes for encoder and decoder hidden layers was set to 500. The dimensionality of latent variables was set to 3. We also implemented the L1 and L2 regularizers to avoid overfitting. To evaluate the performance of the model, the VAE total loss was used to identify the reconstruction error between encoder input and decoder output.

Once the model was trained and the encoder, decoder, and VAE were constructed, the output of the encoder model was selected as the representation of the multidimensional patterns of each county. K-means clustering was used to identify the similar segmentation of the counties. To identify the optimum number of clusters as well as the homogeneity of data points within each cluster, the elbow method [35] and the silhouette score [36] were used.

Explaining the Socioeconomic Characteristics of Similar Counties

To compare the socioeconomic characteristics of the counties in each cluster, the 2016 MIT election data were used as input, while the classes were the cluster labels. The data were divided into training and testing sets with a 70:30 split, respectively. The random forest classifier [37] with 10 k-fold cross-validations was used to build the predictive models. The area under the curve (AUC) of the model was calculated, and the most important features associated with the cluster numbers were defined as the parameters describing the characteristics of counties in each cluster. Feature scores of different census variables for the clusters were computed, which yielded an idea of the relative importance of different socioeconomic and demographic factors for explaining the different clusters. Figure 3 summarizes our approach.

Figure 3. Framework to identify the socioeconomic characteristics of different clusters of counties using random forest feature importance.

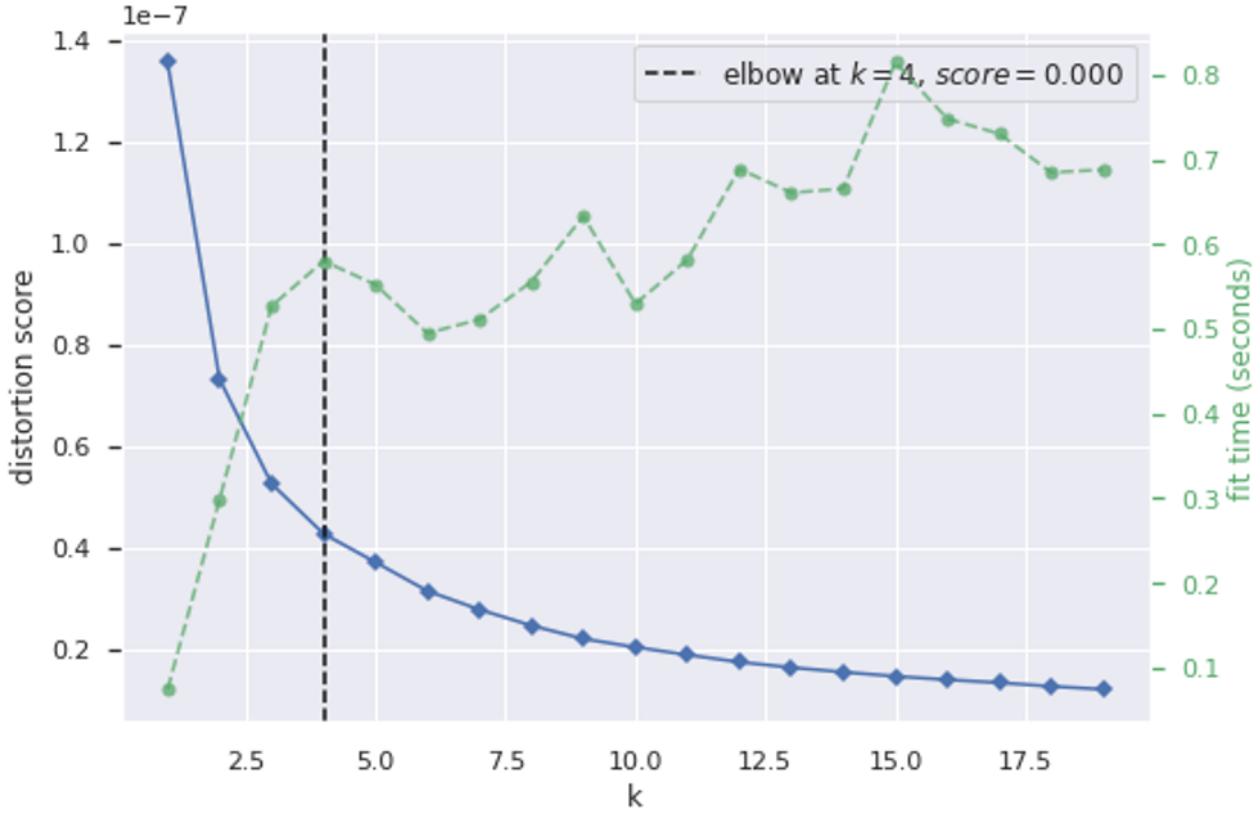


Results

Clustering

This study leveraged a partitioning-based deep learning model to cluster counties based on similarities in social mobility. For each county included in this study, data were clustered based on a combination of the daily values of the 4 Google mobility indicators (retail, grocery and pharmacy, workplace, and residence). The multidimensional time series of Google social mobility indicators from 1089 counties was divided into training and testing sets and fed into the VAE model. The result demonstrated a loss of 0.08. The latent variables were extracted as the output of the encoder. The K-means clustering algorithm identified 4 social mobility clusters. The number of counties in these clusters, which were termed as 0, 1, 2, and 3, were 215, 338, 473, and 59, respectively. Figure 4 gives the distortion scores of the K-means clustering.

Figure 4. Distortion score elbow for K-means clustering.



Google Social Mobility Trends

Across all clusters, visits to retail stores fell significantly after the start of the pandemic until around mid-April, followed by a steady increase and plateauing in early July (Figure 5). Visits to retail outlets began to decline again in late September but then began an upward trend starting on Thanksgiving weekend until the end of December. Retail social mobility values were the highest for cluster 0, followed by clusters 2 and 1, with cluster 3 having the lowest social mobility. Grocery and pharmacy mobility trends reflected those seen for retail social movements but were less pronounced (Figure 6). Cluster 0 had the highest values of grocery mobility, followed by clusters 2, 1, and 3. Workplace mobility showed an initial decline at the start of the pandemic, followed by a steady increase from early May onward (Figure 7). Spikes in mobility were observed during the weekend, which did not significantly decline relative to prepandemic observations. County clusters followed the same order, with cluster 0 having the greatest mobility, followed by clusters 2, 1, and 3.

Figure 5. Google retail mobility.

Figure 6. Google grocery and pharmacy mobility.

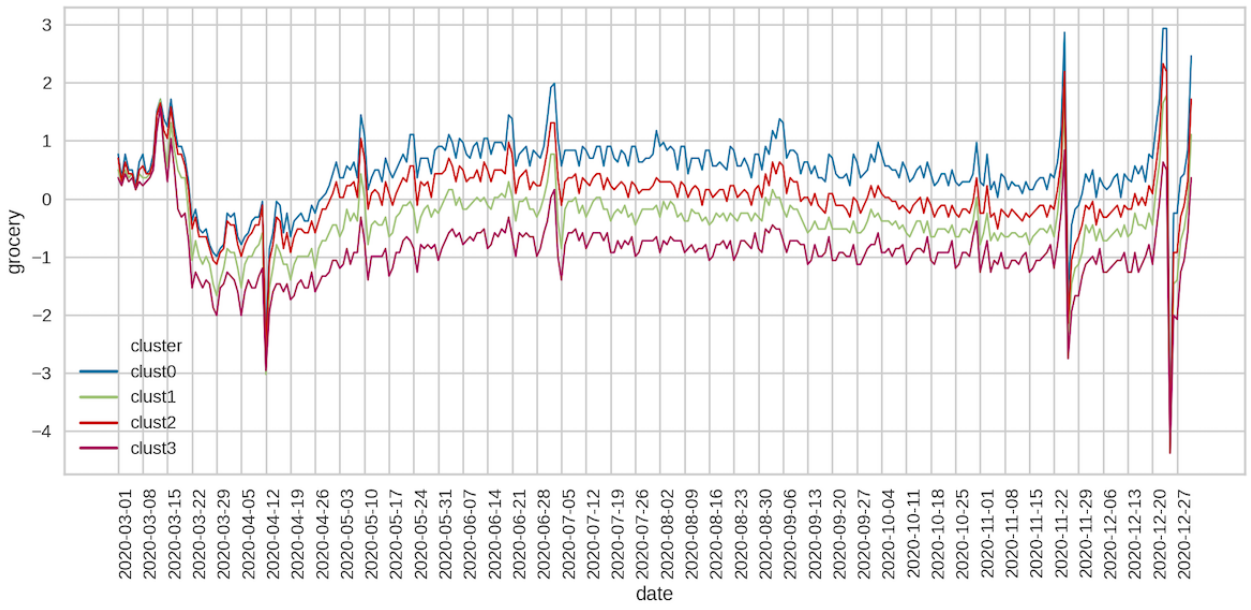
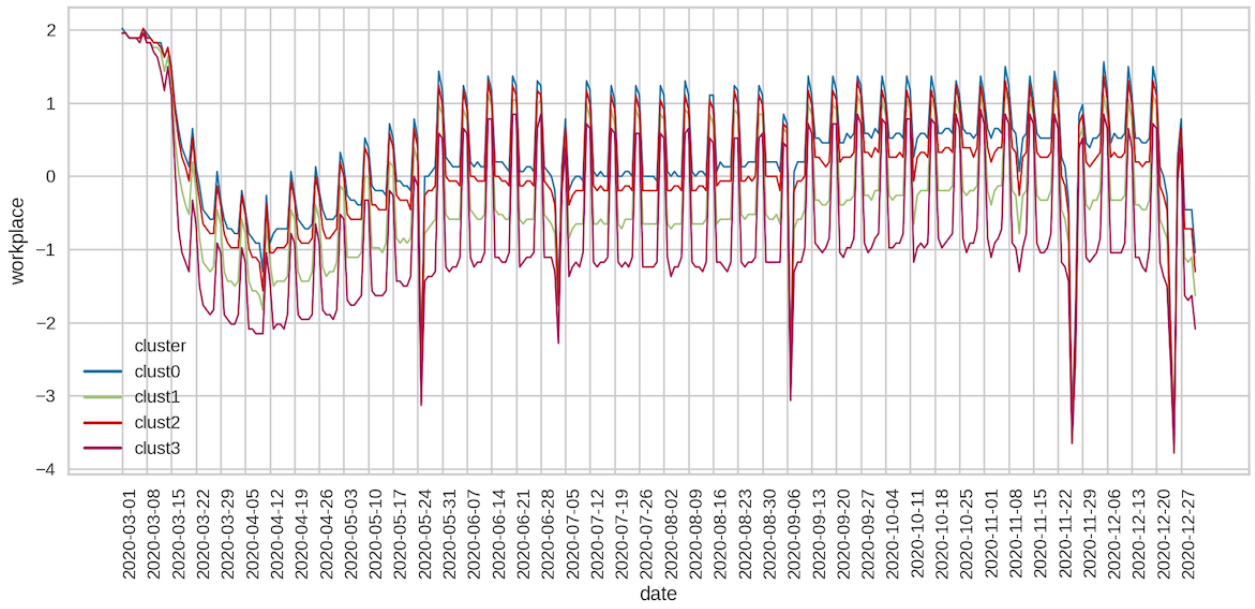
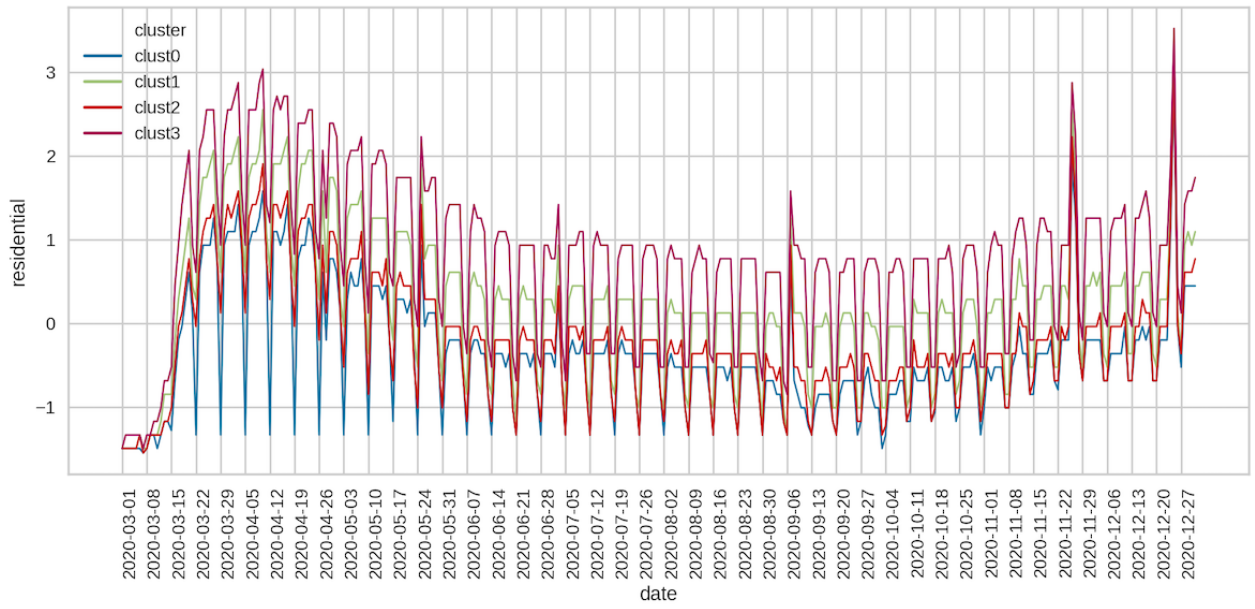


Figure 7. Google workplace mobility.



Finally, residential mobility followed a reverse pattern relative to the other indicators, with cluster 3 having the highest mobility, followed by clusters 1, 2, and 0 (Figure 8). Residential mobility was highest during the onset of the pandemic, followed by a decreasing trend during spring and summer. From late September onward, residential mobility began to increase, and this trend continued until the end of the sample period. The spikes in mobility captured the weekend effects. Our social mobility data indicated differences in mobility between clusters, with counties in cluster 0 having the highest retail, grocery, and workplace mobility and the lowest residential mobility. In contrast, counties in cluster 3 had the lowest social mobility and the highest residential mobility.

Figure 8. Residential mobility.



Relationship Between Google Social Mobility Indicators and County Characteristics

To determine whether county characteristics are correlated with differences in social mobility between the clusters, we obtained socioeconomic, demographic, and political data from each county from 2016 census data [18]. These data included 2016 election returns, race, median income, total population, percentage of rural areas, and education level of the population for age and race. These data were supplemented by county variables collected by other studies [23,25].

A random forest classifier was used to generate feature scores of different socioeconomic and demographic characteristics of the counties included in each cluster, across all 4 clusters (mean

receiver operating characteristic [ROC] AUC 0.871). Table 2 contains the feature scores of all county-level variables.

The top 10 variables in terms of feature scores were percentage of the population aged 65 years and over (0.41715), percentage of females (0.08784), percentage of Whites (0.03869), percentage of Whites with less than college education (0.03772), percentage of Hispanics (0.03369), percentage of Whites with less than high school education (0.03178), percentage of the population using public transit (0.02967), county unemployment rate (0.02759), proportion of voters for Clinton in 2016 (0.02737), and percentage of the population with less than high school population (0.02719). Hence, although political preference and population composition were important, it is important to note the significance of 3 educational variables among the top 10, with the percentage of the population with less than college education being the 11th variable in terms of feature score.

To explore the top 11 socioeconomic, demographic, and political variables impacting social mobility further, we determined the mean population percentage for each county-level variable across clusters (Table 3). The table also contains results of statistical tests of significance of sample means between clusters. The Z test of sample means was performed to compare the significance of different county-level variables for different clusters. Results demonstrated several variable similarities for clusters with the highest social mobility. The percentage of the population aged 65 years and over, Whites, the percentage of whites with less than high school and college education, and the percentage of the overall population with less than college education were higher in counties defined by clusters 0 and 2. Tests of equality of sample proportions and means confirmed that there was a statistically significant difference between clusters 0 and 2 versus clusters 1 and 3 for these population variables. In contrast, the percentage of Hispanics, percentage of the population using public transit for work, and percentage voting for Clinton in 2016 were lower in clusters 0 and 2 relative to clusters 1 and 3. There was no consistent, significant difference across clusters for the percentage of females, population with less than high school education, and unemployment rates.

Table 2. Feature scores of county-level variables.

Feature	Score
Percentage aged 65 years and older	0.41715
Percentage of females	0.08784
Percentage of Whites	0.03869
Percentage of Whites with less than college education	0.03772
Percentage of Hispanics	0.03369
Percentage of Whites with less than high school education	0.03178
Percentage of population using public transit for commuting to work	0.02967
Unemployment rate	0.02759
Percentage voting for Clinton in 2016	0.02737
Percentage with less than high school education	0.02719
Percentage with less than college education	0.02429
Hospitals per 100,000 of population	0.02385
Percentage of rural population	0.0221
Population density	0.02178
Percentage of foreign born	0.02118
Poverty rate	0.02051
Percent without health insurance	0.02003
Percentage voting for Trump in 2016	0.01992
Median household income	0.01911
Percentage aged under 29 years	0.01852
Registered voters as a percentage of population	0.01682
Percentage of African Americans	0.01319

Table 3. Differences in county-level variables across clusters.

Variable	Sample mean (%)				*P* value of sample means between clusters			
	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Clusters 0 and 1	Clusters 0 and 3	Clusters 1 and 2	Clusters 2 and 3
Population aged 65 years and older	17.10	14.20	16.20	13.00	<.01	<.01	<.01	<.01
Females	50.40	50.70	50.70	50.40	.01	.99	.99	.23
White	81.50	66.30	77.10	58.60	<.01	<.01	<.01	<.01
Whites with less than college education	78.20	65.00	75.20	51.10	<.01	<.01	<.01	<.01
Hispanics	6.90	15.50	8.20	19.80	<.01	<.01	<.01	<.01
Whites with less than high school education	11.20	7.10	10.10	5.10	<.01	<.01	<.01	<.01
Population using public transit for commuting to work	0.30	1.20	0.30	3.70	<.01	<.01	<.01	<.01
Unemployment rate	7.50	7.20	7.40	6.30	.06	<.01	.06	<.01
Voting for Clinton in 2016	13.80	20.10	15.30	27.20	<.01	<.01	<.01	<.01
Less than high school education	13.50	11.70	12.60	11.50	<.01	.02	.01	.17
Less than college education	79.50	69.20	76.90	58.50	<.01	<.01	<.01	<.01

Trends in Daily Cases/Deaths by Cluster

Given that policies restricting population mobility were established to curb the spread of COVID-19, we sought to determine whether county clusters with higher social mobility indicators (clusters 0 and 2) reported elevated viral cases and deaths. The daily number of confirmed cases and deaths due to COVID-19 at the county level was obtained from the CSSE at the JHU. We determined the median daily per capita cases (Figure 9) and deaths (Figure 10) by cluster. During the first months of the pandemic, per capita daily cases were quite comparable across clusters (Figure 9). There was a visible divergence that occurred at the beginning of October (onset of the second pandemic wave), with daily cases rising sharply in clusters 0, 1, and 2 relative to cluster 3. For the remainder of the period examined, cluster 0 had the highest number of daily cases, followed by clusters 2 and 1. Cluster 3 retained relatively lower daily cases. Interestingly, clusters 0 and 2 had lower daily deaths until the beginning of September (Figure 10). Daily deaths in these clusters then increased rapidly, and by the beginning of October, per capita deaths in clusters 0, 1, and 2 were higher than in cluster 3.

Figure 9. Daily cases per 100,000 residents.

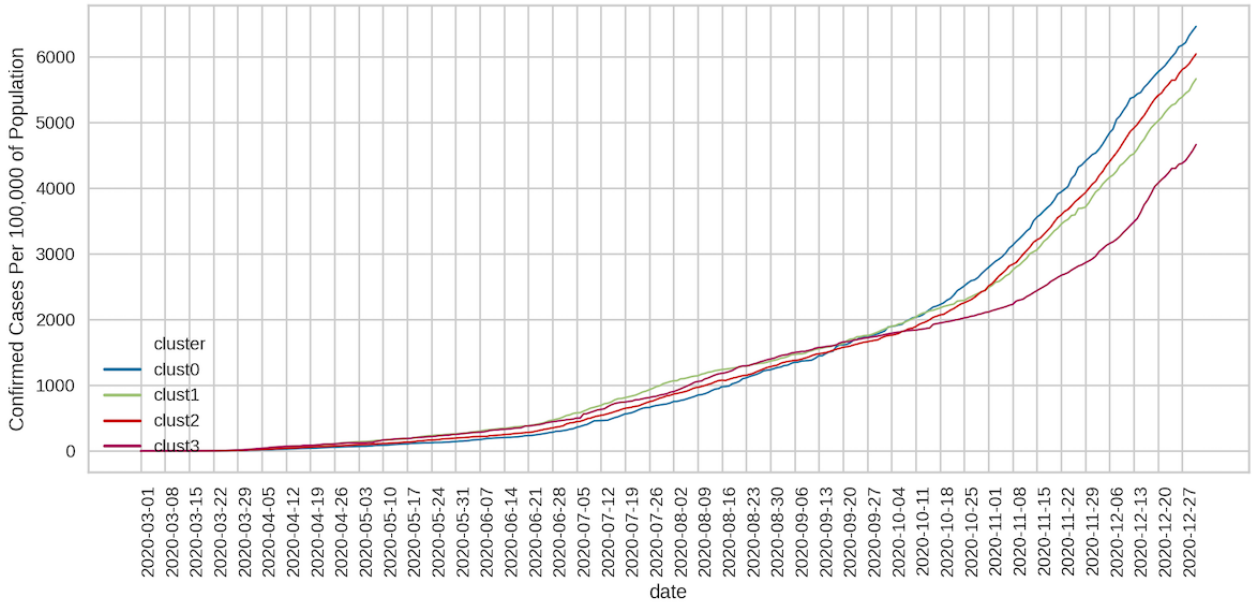
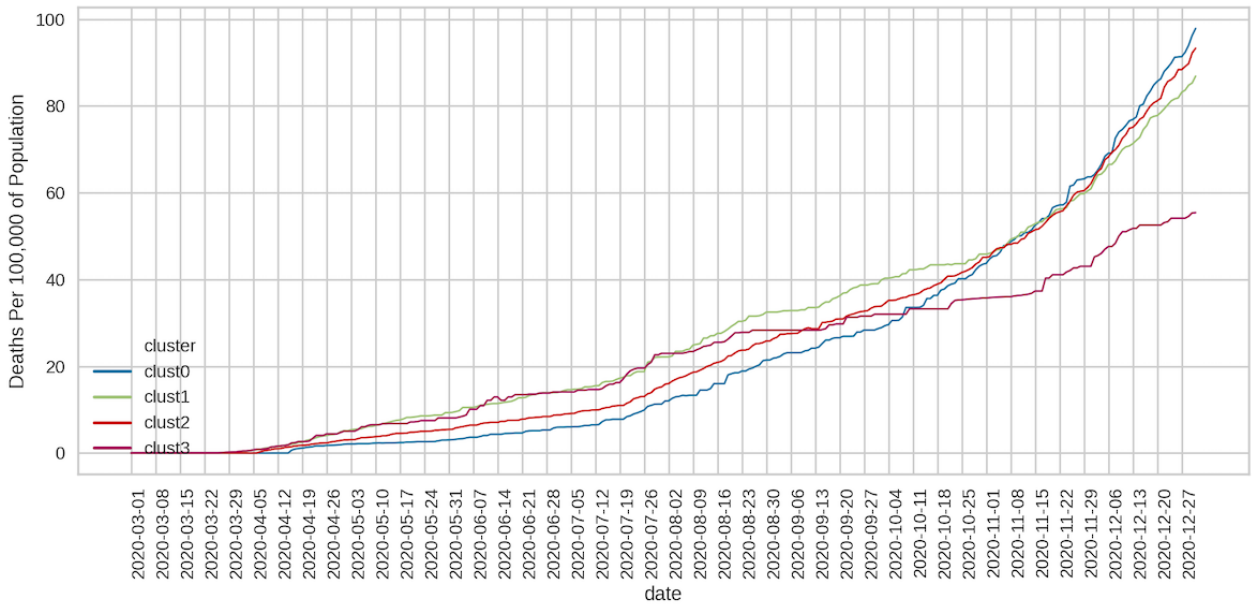


Figure 10. Daily deaths per 100,000 residents.



Discussion

Principal Findings

This study aimed to assess the effect of county-level characteristics on population mobility and the consequences of this mobility on the spread of COVID-19. To the best of our knowledge, this is the first study that has used unsupervised machine learning to understand differences in population mobility across US counties during the first and second waves of the pandemic and determine the relative importance of a wide array of socioeconomic, demographic, and political variables in defining different mobility-based clusters.

Our results demonstrate that of the 4 clusters defined by Google social mobility indicators, the clusters with higher retail, grocery, and work mobility (and lower residential mobility) had several similar population characteristics. Specifically, counties with greater social mobility also had a higher percentage of the population aged 65 years and over, Whites with less than high school and college education, and overall population with less than college education. Counties in these 2 clusters also had a lower share of the population that is Hispanic, the percentage of the population using public transit to work, and the share of voters who voted for Clinton during the 2016 presidential election. Research does suggest that Whites with less than college education constituted a significant voting block for Trump during the 2016 election [38]. In line with this, the 2 clusters with the greatest social mobility also experienced higher per capita COVID-19 case and death rates during most of November and December 2020. These results are consistent with Xie and Li [39], who also used county-level data during the early days of the pandemic and found lower education levels to be correlated with higher infection rates.

The significant increase in COVID-19 cases and deaths in clusters 0 and 2 during November and December 2020 could be a consequence of public rallies and general disregard for social distancing and safety protocols by pro-Trump voters [40]. Although we cannot prove this, the majority of counties in these clusters were Republican leaning during the 2016 presidential election. Moreover, our finding of higher per capita daily COVID-19 cases and deaths in such counties is consistent with other studies. Desmet and Wacziarg [41] found that early on during the pandemic, Republican counties actually experienced lower COVID-19 cases and therefore had lax attitudes toward mask wearing, social distancing, and lockdown measures. However, as the pandemic spread to Trump-leaning counties, population preferences for less stringent social distancing policies had already been formed, making it difficult for policymakers to implement stricter restrictions on social mobility. As a result, this led to greater disease severity in Trump-leaning counties. In a similar vein, Allcott et al [42] found that areas with more Republicans engaged in less social distancing, controlling for other factors, including public policies. In summary, these findings corroborate our own results. Social mobility in the aftermath of the first wave of the pandemic was much higher in Republican counties, which ultimately resulted in higher COVID-19 cases and associated deaths relative to other counties that were Democrat leaning.

Social media is increasingly being used to capture population movements and understand their corresponding impacts on COVID-19 incidence. Social media–based data, including those presented here, have some limitations. Specifically, there is the possibility of sample selection bias if Google Maps users have specific demographic characteristics and are not distributed uniformly across the population. However, data from Statista indicate that in the U.S., Google Maps had 154 million users in April 2018 [43]. Further, published research has done a comparison of Google mobility data against corresponding cellular-generated information by other providers and has found a close correspondence. Specifically, Szocska et al [44] constructed a mobility index and an SAH/resting index based on data on almost all phone subscribers in Hungary and found a close correlation with corresponding Google mobility indices at the national level. There are also a significant number of

published studies that have used Google mobility data to capture population movements for different countries and have found them to be important in predicting movements in COVID-19 (Bryant and Elofsson [11], Askitas et al [45], and Stevens et al [46]). For these reasons, we think there is a high likelihood that Google mobility data do reflect population movements. However, Google mobility data do not include information on certain types of public movements, such as election rallies or community gatherings.

Our research demonstrates the usefulness of publicly available Google mobility data and unsupervised machine learning methods in establishing relationships between county-level characteristics, mobility decisions, and COVID-19 incidence. These findings have important implications for policymakers and public health officials in understanding the effects of NPIs, as the efficacy of such measures on mobility is influenced by underlying socioeconomic, demographic, and political ideology characteristics. The use of Google data enables researchers to assess the types of public movements that are most contributory to COVID-19 spread.

The results of this study provide a unique lens on the potential of machine learning to understand social mobility behaviors. These findings are critical for public health organizations trying to understand the levels of mobility in their counties, in addition to providing insights into some of the underlying factors (ie, social determinants of health) contributing to regional differences in COVID-19 caseloads.

Conclusion

Our results emphasize a role for machine learning methods in public health. Publicly available Google data, in conjunction with census data, can be used to understand the socioeconomic, demographic, and political determinants driving population mobility choices across US counties. This knowledge can assist policymakers in developing NPIs to restrict viral spread during the COVID-19 pandemic.

Acknowledgments

The authors thank Caitlin S Brown, Chris Knittel, and Bora Ozaltun for kindly sharing their data.

Conflicts of Interest

None declared.

References

1. Cucinotta D, Vanelli M. WHO declares COVID-19 a pandemic. Acta Biomed 2020 Mar 19;91(1):157-160 [FREE Full text] [doi: 10.23750/abm.v91i1.9397] [Medline: 32191675]
2. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). URL: https://www.who.int/publications/i/item/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19) [accessed 2022-02-22]
3. Social Distancing. URL: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/social-distancing.html [accessed 2022-02-22]
4. Emergency Preparedness and Response. URL: https://emergency.cdc.gov/ [accessed 2022-02-22]
5. Dave D, Friedson AI, Matsuzawa K, Sabia JJ. When do shelter-in-place orders fight COVID-19 best? Policy heterogeneity across states and adoption time. Econ Inq 2020 Aug 03;59(1):29-52 [FREE Full text] [doi: 10.1111/ecin.12944] [Medline: 32836519]
6. Rosenblatt K. A A Summer of Digital Protest: How 2020 Became the Summer of Activism Both Online and Offline. URL: https://www.nbcnews.com/news/us-news/summer-digital-protest-how-2020-became-summer-activism-both-online-n1241001 [accessed 2022-02-22]
7. See Reopening Plans and Mask Mandates for All 50 States. URL: https://www.nytimes.com/interactive/2020/us/states-reopen-map-coronavirus.html [accessed 2022-02-22]
8. Xiong C, Hu S, Yang M, Luo W, Zhang L. Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. Proc Natl Acad Sci U S A 2020 Nov 03;117(44):27087-27089 [FREE Full text] [doi: 10.1073/pnas.2010836117] [Medline: 33060300]
9. Bisanzio D, Kraemer MUG, Bogoch II, Brewer T, Brownstein JS, Reithinger R. Use of Twitter social media activity as a proxy for human mobility to predict the spatiotemporal spread of COVID-19 at global scale. Geospat Health 2020 Jun 15;15(1):1-17 [FREE Full text] [doi: 10.4081/gh.2020.882] [Medline: 32575957]
10. Li Z, Li X, Porter D, Zhang J, Jiang Y, Olatosi B, et al. Monitoring the spatial spread of COVID-19 and effectiveness of control measures through human movement data: proposal for a predictive model using big data analytics. JMIR Res Protoc 2020 Dec 18;9(12):e24432 [FREE Full text] [doi: 10.2196/24432] [Medline: 33301418]
11. Bryant P, Elofsson A. Estimating the impact of mobility patterns on COVID-19 infection rates in 11 European countries. PeerJ 2020;8:e9879 [FREE Full text] [doi: 10.7717/peerj.9879] [Medline: 32983643]
12. Sulyok M, Walker M. Community movement and COVID-19: a global study using Google's Community Mobility Reports. Epidemiol Infect 2020 Nov 13;148:e284 [FREE Full text] [doi: 10.1017/S0950268820002757] [Medline: 33183366]
13. Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C. Geo-located Twitter as proxy for global mobility patterns. Cartogr Geogr Inf Sci 2014 May 27;41(3):260-271 [FREE Full text] [doi: 10.1080/15230406.2014.890072] [Medline: 27019645]

14. Wang HY, Yamamoto N. Using a partial differential equation with Google Mobility data to predict COVID-19 in Arizona. Math Biosci Eng 2020 Jul 13;17(5):4891-4904 [FREE Full text] [doi: 10.3934/mbe.2020266] [Medline: 33120533]
15. Badr HS, Du H, Marshall M, Dong E, Squire MM, Gardner LM. Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. Lancet Infect Dis 2020 Nov;20(11):1247-1254 [FREE Full text] [doi: 10.1016/S1473-3099(20)30553-3] [Medline: 32621869]
16. Gatalo O, Tseng K, Hamilton A, Lin G, Klein E, CDC MInD-Healthcare Program. Associations between phone mobility data and COVID-19 cases. Lancet Infect Dis 2021 May;21(5):e111 [FREE Full text] [doi: 10.1016/S1473-3099(20)30725-8] [Medline: 32946835]
17. Karaivanov A, Lu S, Shigeoka H, Chen C, Pamplona S. Face masks, public policies and slowing the spread of COVID-19: Evidence from Canada. J Health Econ 2021 Jul;78:102475 [FREE Full text] [doi: 10.1016/j.jhealeco.2021.102475] [Medline: 34157513]
18. Wang S, Liu Y, Hu T. Examining the change of human mobility adherent to social restriction policies and its effect on COVID-19 cases in Australia. Int J Environ Res Public Health 2020 Oct 29;17(21):7930 [FREE Full text] [doi: 10.3390/ijerph17217930] [Medline: 33137958]
19. Glaeser E, Gorback C, Redding S. JUE insight: how much does COVID-19 increase with mobility? Evidence from New York and four other U.S. cities. J Urban Econ 2020 Oct 21:103292 [FREE Full text] [doi: 10.1016/j.jue.2020.103292] [Medline: 33106711]
20. Gao S, Rao J, Kang Y, Liang Y, Kruse J, Dopfer D, et al. Association of mobile phone location data indications of travel and stay-at-home mandates with COVID-19 infection rates in the US. JAMA Netw Open 2020 Sep 01;3(9):e2020485 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.20485] [Medline: 32897373]
21. Narayanan RP, Nordlund J, Pace RK, Ratnadiwakara D. Demographic, jurisdictional, and spatial effects on social distancing in the United States during the COVID-19 pandemic. PLoS One 2020;15(9):e0239572 [FREE Full text] [doi: 10.1371/journal.pone.0239572] [Medline: 32960932]
22. Grossman G, Kim S, Rexer JM, Thirumurthy H. Political partisanship influences behavioral responses to governors' recommendations for COVID-19 prevention in the United States. Proc Natl Acad Sci U S A 2020 Sep 29;117(39):24144-24153 [FREE Full text] [doi: 10.1073/pnas.2007835117] [Medline: 32934147]
23. Knittel C, Ozaltun B. What Does and Does Not Correlate with COVID-19 Death Rates (Report No.: w27391). 2020. URL: https://www.nber.org/papers/w27391 [accessed 2022-02-22]
24. McLaren J. Racial Disparity in COVID-19 Deaths: Seeking Economic Roots with Census Data (Report No.: w27407). URL: https://www.nber.org/papers/w27407 [accessed 2022-02-22]
25. Brown C, Ravallion M. Inequality and the Coronavirus: Socioeconomic Covariates of Behavioral Responses and Viral Outcomes Across US Counties (Report No.: w27549). URL: https://www.nber.org/papers/w27549 [accessed 2022-02-22]
26. Center for Systems Science and Engineering (CSSE). COVID-19 Dashboard. URL: https://coronavirus.jhu.edu/map.html [accessed 2022-02-22]
27. MEDSL/2018-elections-unoffical. URL: https://github.com/MEDSL/2018-elections-unoffical [accessed 2022-02-22]
28. Song C, Liu F, Huang Y, Wang L, Tan T. Auto-encoder Based Data Clustering. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Berlin; Heidelberg: Springer; 2013:117-124.
29. Nagano M, Nakamura T, Nagai T, Mochihashi D, Kobayashi I, Takano W. HVGH: unsupervised segmentation for high-dimensional time series using deep neural compression and statistical generative model. Front Robot AI 2019;6:115 [FREE Full text] [doi: 10.3389/frobt.2019.00115] [Medline: 33501130]
30. Jalali N, Sahu K, Oetomo A, Morita P. Understanding user behavior through the use of unsupervised anomaly detection: proof of concept using Internet of Things smart home thermostat data for improving public health surveillance. JMIR Mhealth Uhealth 2020 Nov 13;8(11):e21209 [FREE Full text] [doi: 10.2196/21209] [Medline: 33185562]
31. Chung NC, Mirza B, Choi H, Wang J, Wang D, Ping P, et al. Unsupervised classification of multi-omics data during cardiac remodeling using deep learning. Methods 2019 Aug 15;166:66-73 [FREE Full text] [doi: 10.1016/j.ymeth.2019.03.004] [Medline: 30853547]
32. Variational Recurrent Autoencoder for Timeseries Clustering in Pytorch. URL: https://pythonawesome.com/variational-recurrent-autoencoder-for-timeseries-clustering-in-pytorch/ [accessed 2022-02-22]
33. Yu X, Li H, Zhang Z, Gan C. The optimally designed variational autoencoder networks for clustering and recovery of incomplete multimedia data. Sensors (Basel) 2019 Feb 16;19(4):809 [FREE Full text] [doi: 10.3390/s19040809] [Medline: 30781499]
34. Keras Team. Keras: The Python Deep Learning API. URL: https://keras.io/ [accessed 2022-02-22]
35. Using the Elbow Method to Determine the Optimal Number of Clusters for K-means Clustering. URL: https://bl.ocks.org/rpgove/0060ff3b656618e9136b [accessed 2022-02-22]
36. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987 Nov;20:53-65. [doi: 10.1016/0377-0427(87)90125-7]
37. Rogers JD. Identifying Feature Relevance using a Random Forest. URL: http://videolectures.net/slsfs05_rogers_ifrur/ [accessed 2022-02-22]

38. Jones B. An Examination of the 2016 Electorate, Based on Validated Voters. 2018. URL: https://www.pewresearch.org/politics/2018/08/09/an-examination-of-the-2016-electorate-based-on-validated-voters/ [accessed 2022-02-22]
39. Xie Z, Li D. Health and Demographic Impact on COVID-19 Infection and Mortality in US Counties. URL: https://www.medrxiv.org/content/10.1101/2020.05.06.20093195v1 [accessed 2022-02-22]
40. Sanchez B. Trump Supporter on Not Wearing a Mask: It's a Fake Pandemic. URL: https://www.cnn.com/videos/politics/2020/09/11/trump-rally-attendees-michigan-ctn-vpx.cnn [accessed 2022-02-22]
41. Desmet K, Wacziarg R. Understanding spatial variation in COVID-19 across the United States. J Urban Econ 2021 Mar 11:103332 [FREE Full text] [doi: 10.1016/j.jue.2021.103332] [Medline: 33723466]
42. Allcott H, Boxell L, Conway J, Gentzkow M, Thaler M, Yang D. Polarization and public health: partisan differences in social distancing during the coronavirus pandemic. J Public Econ 2020 Nov;191:104254 [FREE Full text] [doi: 10.1016/j.jpubeco.2020.104254] [Medline: 32836504]
43. Most Popular Social Media Apps in U.S. URL: https://www.statista.com/statistics/248074/most-popular-us-social-networking-apps-ranked-by-audience/ [accessed 2022-02-22]
44. Szocska M, Pollner P, Schiszler I, Joo T, Palicz T, McKee M, et al. Countrywide population movement monitoring using mobile devices generated (big) data during the COVID-19 crisis. Sci Rep 2021 Mar 15;11(1):5943 [FREE Full text] [doi: 10.1038/s41598-021-81873-6] [Medline: 33723282]
45. Askitas N, Tatsiramos K, Verheyden B. Estimating worldwide effects of non-pharmaceutical interventions on COVID-19 incidence and population mobility patterns using a multiple-event study. Sci Rep 2021 Jan 21;11(1):1972 [FREE Full text] [doi: 10.1038/s41598-021-81442-x] [Medline: 33479325]
46. Stevens NT, Sen A, Kiwon F, Morita PP, Steiner SH, Zhang Q. Estimating the effects of non-pharmaceutical interventions and population mobility on daily COVID-19 cases: evidence from Ontario. Can Public Policy 2021 Dec 24:e2021022 [FREE Full text] [doi: 10.3138/cpp.2021-022]

Abbreviations

AUC: area under the curve
CSSE: Center for Systems Science and Engineering
JHU: Johns Hopkins University
LSTM: long short-term memory
MIT: Massachusetts Institute of Technology
NPI: nonpharmaceutical intervention
SAH: stay-at-home
SIP: shelter-in-place
VAE: variational autoencoder

Edited by R Cuomo; submitted 06.07.21; peer-reviewed by J Wang, Z Xie; comments to author 07.10.21; revised version received 26.12.21; accepted 13.01.22; published 03.03.22

Please cite as:
Jalali N, Tran NK, Sen A, Morita PP
Identifying the Socioeconomic, Demographic, and Political Determinants of Social Mobility and Their Effects on COVID-19 Cases and Deaths: Evidence From US Counties
JMIR Infodemiology 2022;2(1):e31813
URL: https://infodemiology.jmir.org/2022/1/e31813
doi: 10.2196/31813
PMID: 35287305