Identifying the Socioeconomic, Demographic, and Political Determinants of Social Mobility and their Effects on COVID-19 Cases and Deaths: Evidence from U.S. Counties

Niloofar Jalali, N. Ken Tran, Anindya Sen, Plinio Pelegrini Morita

Submitted to: JMIR Infodemiology
on: July 06, 2021

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Identifying the Socioeconomic, Demographic, and Political Determinants of Social Mobility and their Effects on COVID-19 Cases and Deaths: Evidence from U.S. Counties

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

[1]University of Waterloo School of Public Health and Health Systems, Faculty of Applied Health Sciences, Waterloo CA
[2]University of Waterloo Department of economics Waterloo CA
[3]University of Waterloo Acting Associate Dean (Co-operative Education & Planning) Faculty of Arts Waterloo CA
[4]University of Waterloo School of Public Health and Health Systems Waterloo CA
[5]Department of Systems Design Engineering, University of Waterloo Waterloo CA

Corresponding Author:
Plinio Pelegrini Morita MSc, PEng, PhD
University of Waterloo
School of Public Health and Health Systems
200 University Avenue West Waterloo
Waterloo
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 non-pharmaceutical interventions to control the spread of COVID-19 depend on our understanding of the determinants of public mobility.

Objective: This study used social media data and machine learning to investigate population mobility across a sample of U.S. 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 1,085 U.S. counties from March 1st, 2020 to December 31st, 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 U.S. counties into four mobility clusters. Clusters with higher population mobility had a higher percentage of the population aged 65 and over, a higher percentage of Whites with less than high school and college education, a larger percentage of the population with less than a college education, a smaller share of the population that is Hispanic, 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, those clusters with greater social mobility experienced a sharp increase in per capita COVID-19 case and death rates from October to December 2020.

Conclusions: These results emphasize the importance of using Google data and machine learning methods in public health data to support the identification of underlying determinants of social mobility patterns and associated COVID-19 cases.

(JMIR Preprints 06/07/2021:31813)
DOI: https://doi.org/10.2196/preprints.31813

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Identifying the Socioeconomic, Demographic, and Political Determinants of Social Mobility and their Effects on COVID-19 Cases and Deaths: Evidence from U.S. Counties

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 non-pharmaceutical interventions to control the spread of COVID-19 depends 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 U.S. 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 1,085 U.S. counties from March 1st, 2020 to December 31st, 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 U.S. counties into four Google mobility clusters. Clusters with more population mobility had a higher percentage of the population aged 65 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 - 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.

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 around the world 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 inter-personal physical contact or reducing the movement of the population was an effective means to control the spread of the virus [2]. These findings spurred national and sub-national 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].

Besides the direct impacts of such policies, evaluating the effects of demographic and socioeconomic factors on population mobility are 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 United States [6]. Increases in social mobility also occurred as many jurisdictions modified their stay-at-home orders, allowed more businesses to re-open, 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 U.S. Presidential Elections.

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 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 for 25 counties provided by Teralytics and found reduced mobility patterns were 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) was associated with a reduced spread of COVID-19 across states. Glaeser et al.[19] also employ Safegraph data and found reduced mobility to be correlated with lower cases for some U.S. cities. Employing Google data for 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 U.S. counties during the early days of the pandemic. As such, they were unable to capture socioeconomic, demographic, and political determinants of mobility[22–26].

We evaluate the determinants and consequences of population movements in 1,089 U.S. 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 U.S. counties. We also investigate the effects of such mobility decisions on daily per capita COVID-19 cases and deaths. Social mobility is 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 non-pharmaceutical interventions (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 lower duration of stay at residence also had a higher percentage of population aged 65 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-December 2020. These findings are consistent with other studies, which 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.

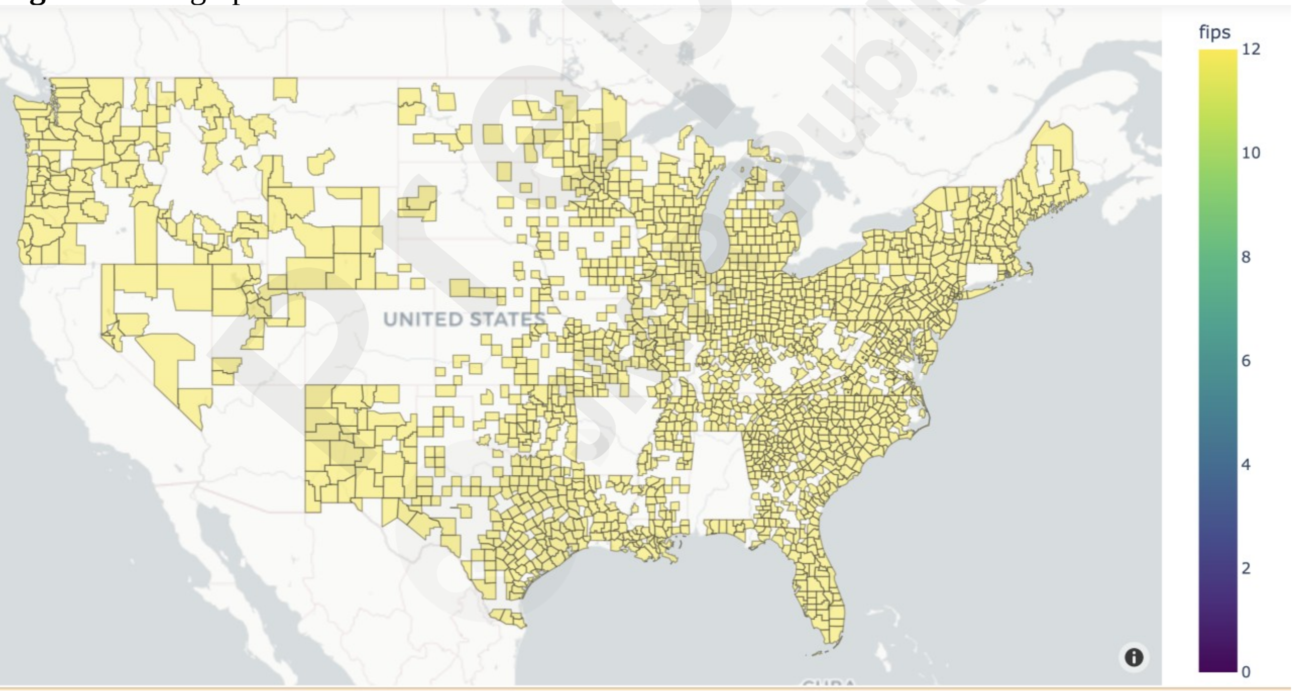
METHODS

Data

COVID-19 Incidence Data

The daily number 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) [27]. For the 1,089 counties in our sample, the mean (standard deviation) of confirmed cases and deaths (both per 100,000 of the population) are 1,541.27 (1905.59) and 33.72 (44.78), respectively. Figure 1 below 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 below) 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 was 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 of subway, bus, and train stations; retail stores & recreation outlets consisting of places like restaurants, cafes, shopping centers, theme parks, museums, libraries, and movie theaters; and workplaces. The Google Mobility data also provides 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 pre-pandemic baseline mobility value was determined using the median mobility for each day of the week from January 3rd to February 6th, 2020 [28]. Subsequent mobility values were normalized to the 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 three prior days. The availability of Google data determined which counties we used in our analysis. The final dataset contains observations for 1,089 counties, which is roughly 35% of the total number of counties (3,142) in the United States. Daily values are available for the first and second waves of the pandemic from March 11th, 2020 to December 31st, 2020

With the exception of the residential index, daily values for each index are calculated relative to a baseline, which is defined as the median for the corresponding day of the week, during the 5-week period 3 January – 6 February 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 above 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

2016 census data was collected by the MIT Election Data and Science Lab [18]. These data have been supplemented by county variables collected by other studies [24,26]. To validate that our samples were representative of all U.S. counties, we compiled summary statistics of socio-economic and demographic variables between our sample and all counties (Table 1). In summary, there does not seem to be dramatic differences in most variables between all counties and our sample. The exception is population, where our sample mean is more than 2.5 times that of the mean for all counties. In a similar vein, while all counties have 58% of the population in rural areas, the corresponding statistic for our sample is 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

	All Counties				Our Sample			
Variable	**Mean**	**Std Dev**	**Min**	**Max**	**Mean**	**Std Dev**	**Min**	**Max**
Politics								
Population voting for Trump 2016 (%)	28.13	8.44i	1.93	76.32	24.28	7.22	2.63	66.42
Population voting for Clinton 2016 (%)	14.07	7.41	0.00	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	0.00	86.19	9.96	12.21	0.09	76.55
Hispanics (%)	8.99	13.66	0.00	98.96	11.03	13.41	0.68	95.48
Foreign born (%)	4.62	5.63	0.00	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 and Under (%)	37.34	5.44	11.84	70.98	39.24	4.98	13.64	61.69
Population aged 65 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	11.04	5.33	0.00	41.76	9.11	3.92	0.97	25.57

less than high school education (%)								
Whites with less than college education (%)	77.00	10.36	9.19	95.92	71.28	11.58	15.30	89.96
Socioeconomic								
Median household income	47817.6	12482.4	18972	125672.	53798.5	13905.9	28452	125672
Rural population (%)	58.48	31.45	0.00	100.00	31.733	22.08	0.00	100.00
Population density	582.71	3761.83	0.26	179922.3	1397.32	6127.90	6.22	179922.3
Hospitals per 100,000 of population	0.61	0.94	0.00	10.56	0.25	0.166	0.00	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 (%)	0.00	0.01	0.00	0.26	0.01	0.02	0.00	0.26

Clustering

Figure 2. Methodology for identifying different clusters of counties using Variational Autoencoder

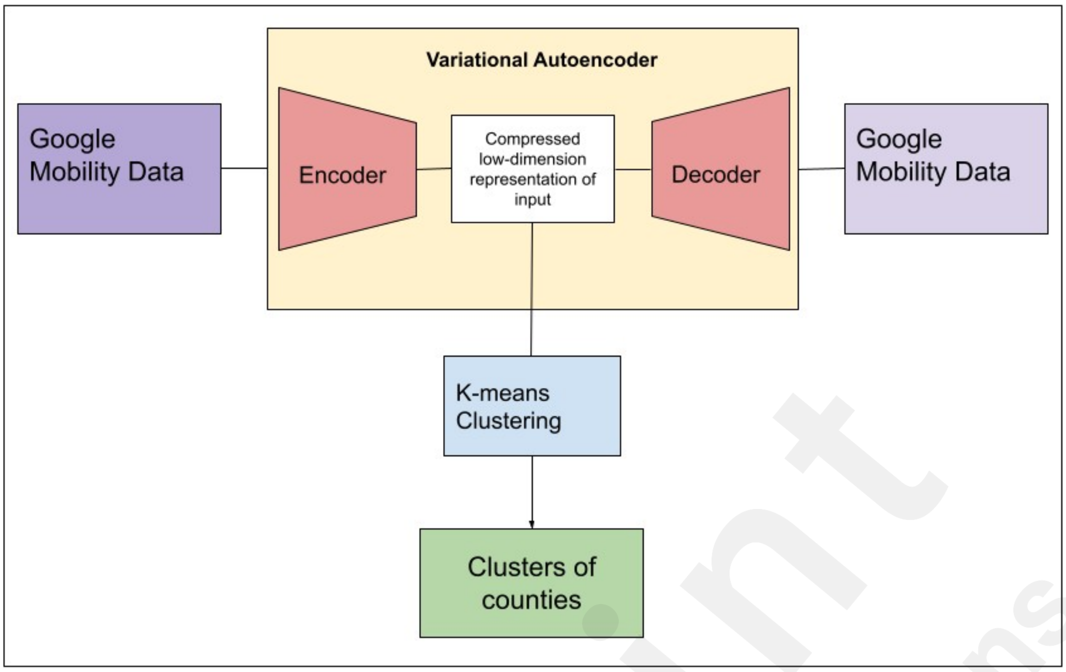


Figure 2 summarizes our methodology in identifying different clusters of counties using Google mobility indicators. Clustering is an unsupervised learning technique that partitions the dataset into groups or clusters based on similarity measures. This study leveraged partitioning-based algorithms, which divided the dataset 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 four Google mobility indicators. To identify the different clusters of counties, we performed two steps [29];

1. Compressing the multi-dimensional 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

To compress the multi-dimensional time series, we implemented the Variational Autoencoder (VAE) architecture based on long short-term memory (LSTM) [30–32]. The principal concept of this generative approach is to project the high-dimensional data into latent variables. Our model comprises four blocks [33];

- 1. Encoder: Defined by the LSTM layers, the multi-dimensional time-series input (x) are fed to the LSTM.
- 2. Encoder to latent layer: Defined by a linear layer, that identifies the mean and standard deviations of the last hidden layer of the encoder. During the training process, the multi-gaussian distributions are defined and reparametrized iteratively, by the mean and standard deviations derived from latent vectors.
- 3. Latent-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, that uses latent variables (z) to reconstruct the original data [34].

Identifying the true posterior distribution is intractable [34]. Therefore, to construct the original data, the probabilistic encoder model was approximated by normal distribution $p(z \vee x) N(0,1)$ and used as a probability decoder [31,34]. 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:

1. The divergence from the approximated distribution and the true distribution

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

$$D_{KL}[N(\mu_x, \sigma_x^2) \vee \text{¿} N(0,1)] = \frac{1}{2} \sum_k \left(\exp\left(\sigma_x^2\right) + \mu_x^2 - 1 - \sigma_x^2\right) \qquad (2)$$

2.The Mean squared error loss calculated the difference between original and reconstructed input data. ($mse(x - \hat{x})^2$)
3. The total loss is defined as sum of two losses:

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

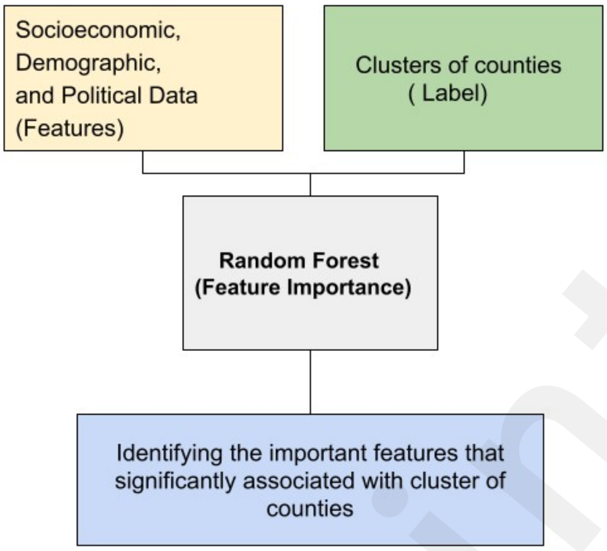
The model was trained in Python 3.6 using the Keras library [35] with *Adam* optimizer. The batch size and number of the epochs were set to 10 and 100 respectively. The number of nodes for encoder and decoder hidden layers was set to 500. The dimensionality of latent variables was set to 3. We also implemented the *L1* and *L2* regularizer to avoid overfitting. To evaluate the performance of the model, the VAE total loss is used to identify the reconstruction error between encoder input and the decoder output.

Once the model was trained and the encoder, decoder, and *VAE* were constructed, the output of the encoder model is selected as the representation of multi-dimensional patterns of each county. The *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 [36] and silhouette score [37] were used.

Explaining the Socio-economic Characteristics of Similar Counties

To compare the socio-economic characteristics of the counties in each cluster, the MIT election (2016) 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 [38] with 10 k-fold cross-validations was used to build the predictive models. The area under 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

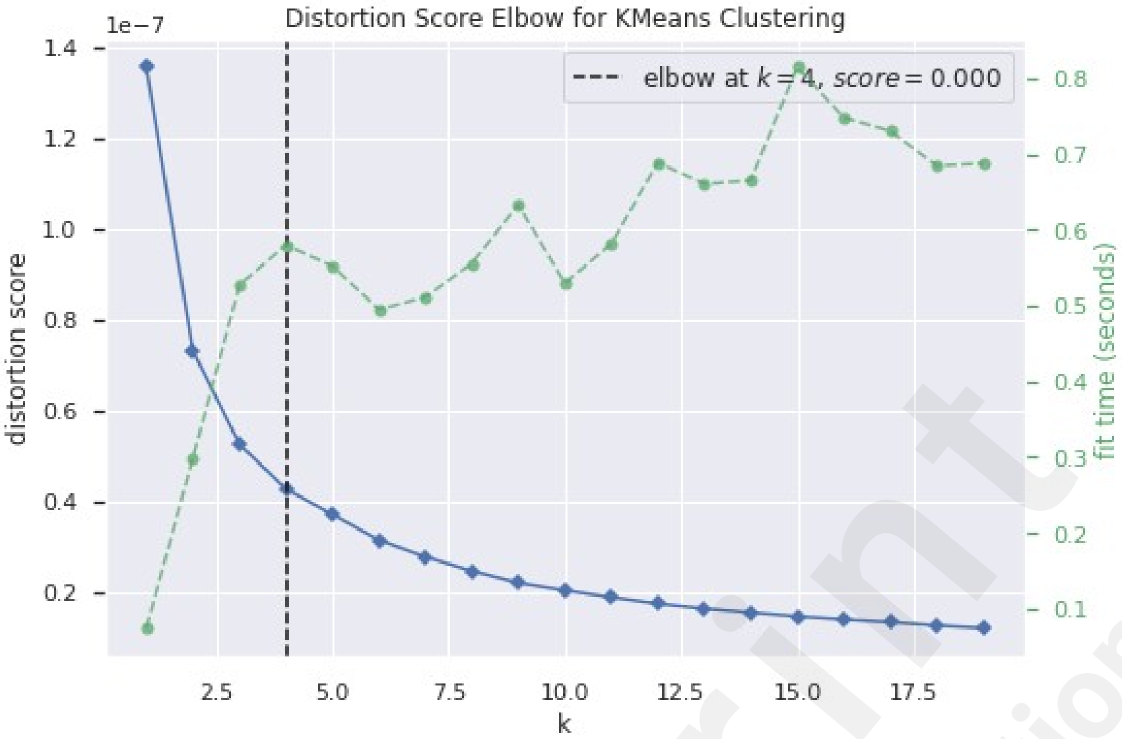

Socioeconomic, Demographic, and Political Data (Features)
Clusters of counties (Label)
Random Forest (Feature Importance)
Identifying the important features that significantly associated with cluster of counties

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 four Google mobility indicators (retail, grocery and pharmacy, workplace, and residence). The multi-dimensional time series of Google social mobility indicators from 1,089 counties was divided into training and testing sets and fed into the VAE model. The result demonstrated the loss of 0.08. The latent variables are extracted as the output of the encoder. The K-means clustering algorithm identified four social mobility clusters. The number of counties in these clusters – which are termed as 0, 1, 2, and 3 - are 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 upwards trend starting on Thanksgiving Weekend until the end of December. Retail social mobility values are 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 onwards (Figure 7). Spikes in mobility were observed during the weekend, which did not significantly decline relative to pre-pandemic observations. County clusters follow the same order with cluster 0 having the greatest mobility followed by clusters 2, 1, and 3.

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 onwards, residential mobility began to increase, and this trend continued till the end of the sample period. The spikes in mobility capture 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 highest residential mobility.

Figure 5. Google Retail Mobility

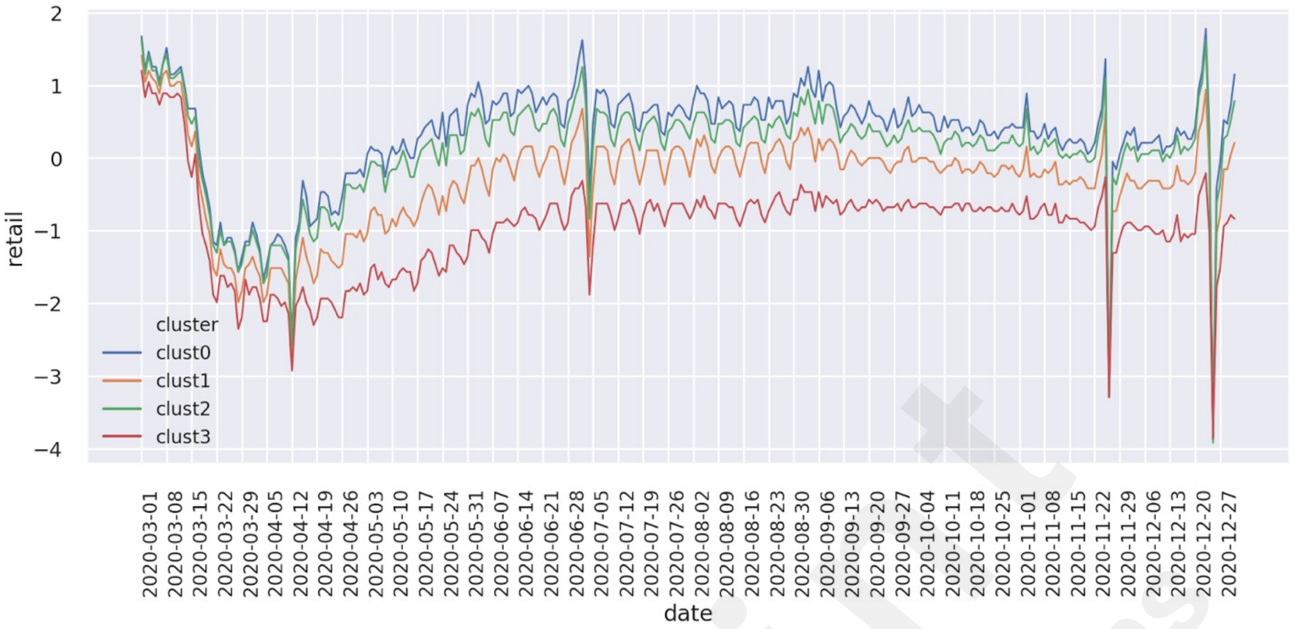


Figure 6. Google Groceries & Pharmacies Mobility

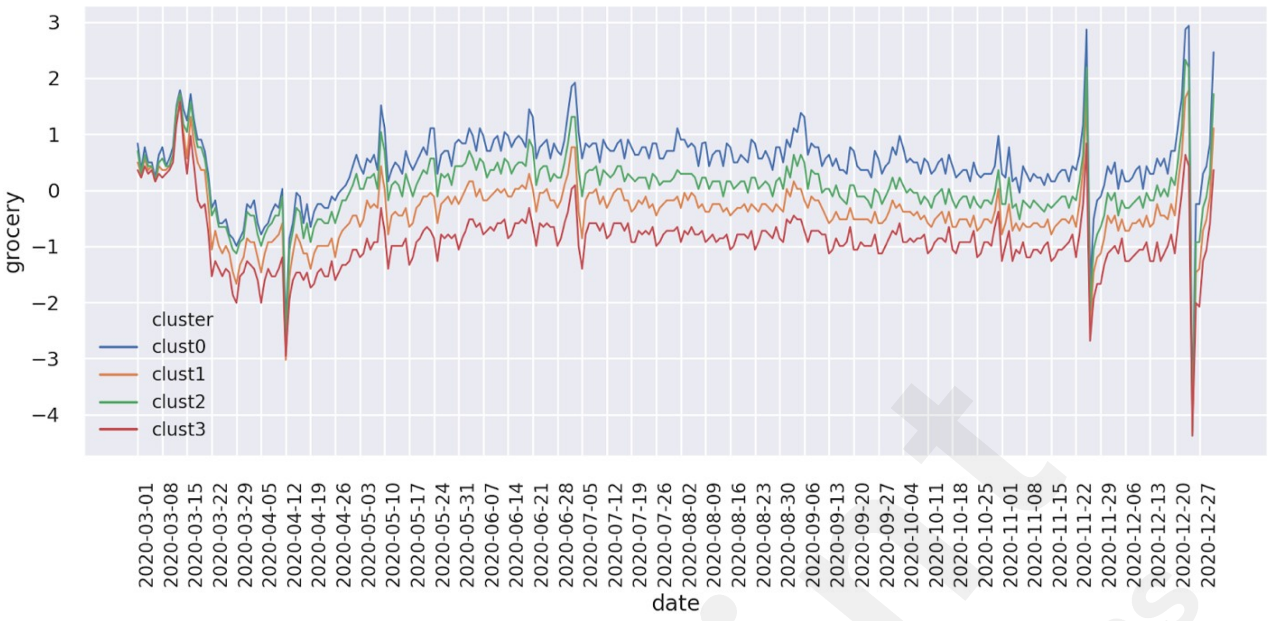


Figure 7. Google Workplace Mobility

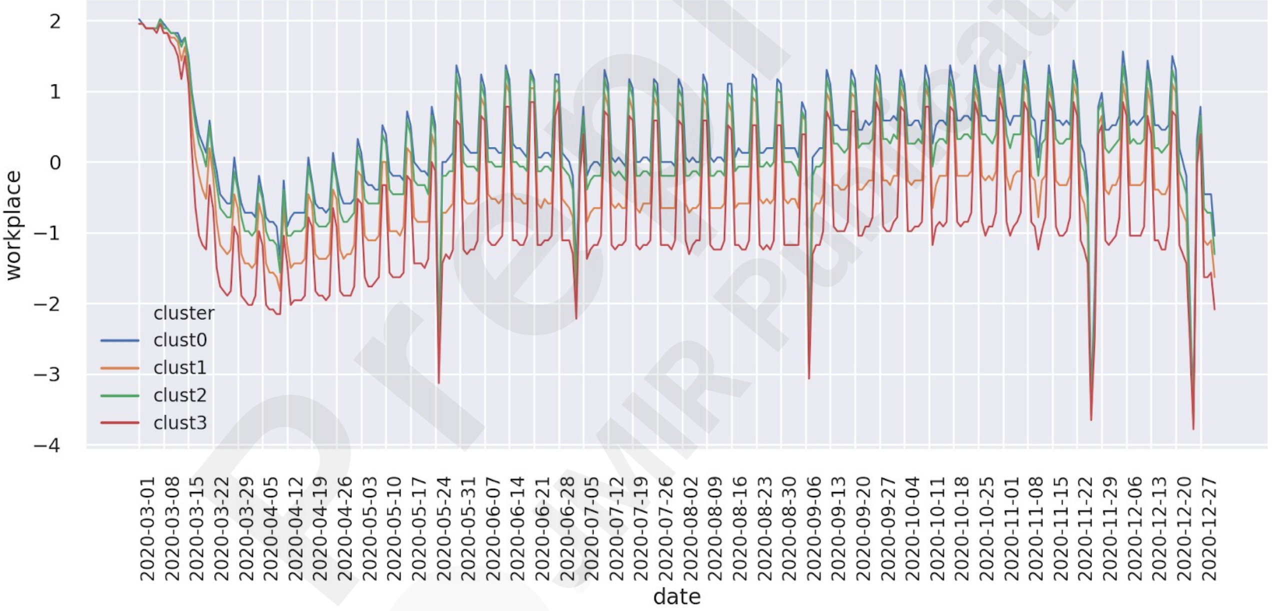
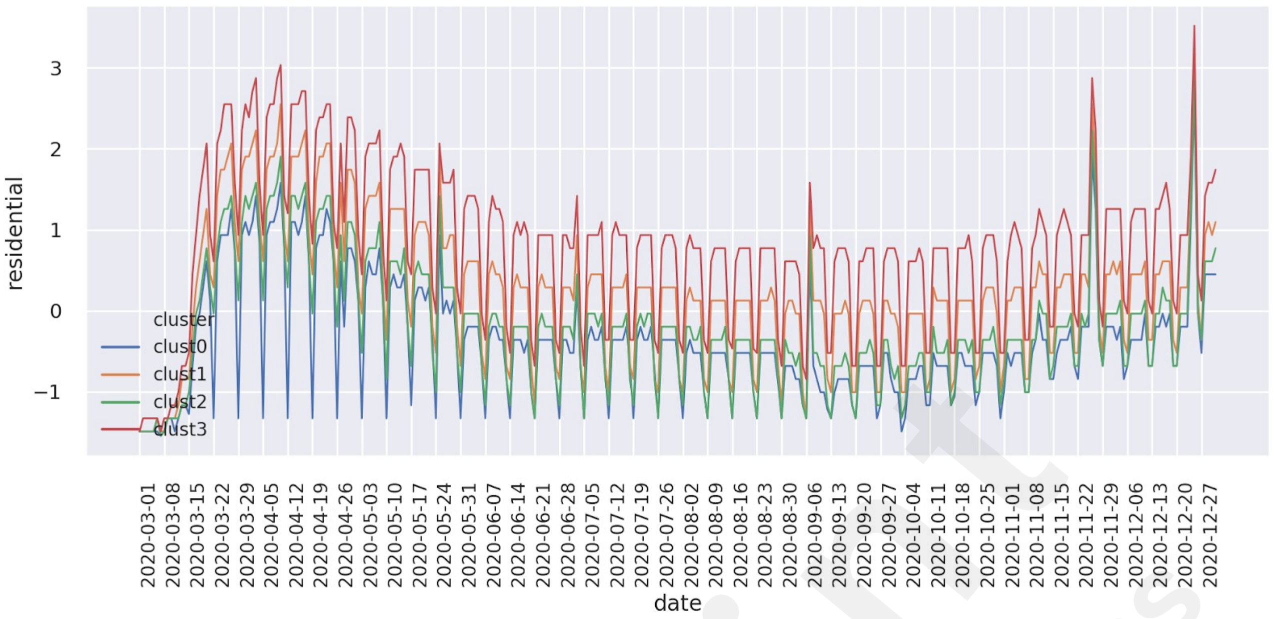


Figure 8. Residential Mobility



Relationship between Google Social Mobility Indicators and County Characteristics

To determine whether county characteristics were correlated with differences in social mobility between the clusters, we obtained socioeconomic, demographic, and political data from each county from 2016 census data [18]. This data includes 2016 election returns, race, median income, total population, the percentage of rural areas, and the education level of the population for age and race. These data have been supplemented by county variables collected by other studies [24,26].

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 four clusters (Mean ROC AUC: 0.871). Table 2 contains the feature scores of all county-level variables.

Table 2. Feature Scores of County Level Variables

Feature	**Score**
Percentage Aged 65 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
Share of Population Using Public Transit for Commuting to Work (%)	0.02967
Unemployment Rate	0.02759
Percentage Voting for Clinton 2016	0.02737
Percentage less than High School Education	0.02719
Percentage less than College Education	0.02429
Hospitals Per 100,000 of Population	0.02385
Percentage 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 2016	0.01992
Median Household Income	0.01911
Percentage Aged and Under 29	0.01852
Registered Voters as Percentage of Population	0.01682
Percentage of African Americans	0.01319

The top 10 variables in terms of feature scores were: percentage of the population aged 65 and over (0.42); percentage of females (0.088), percentage of Whites (0.039); percentage of Whites with less than a college education (0.038); percentage of Hispanics (0.03); percentage of Whites with less than high school education (0.03); share of the population using public transit (0.03); county unemployment rate (0.027); the proportion of voters for Clinton 2016 (0.027); and percentage of the population with less than high school population (0.027). Hence, while political preference and population composition were important, it is important to note the significance of three educational variables among the top 10, with the percentage of the population with less than college education being the eleventh variable in terms of feature score.

To explore the top eleven 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. Z test of sample means were 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 Population Aged 65 and

Over, Whites, the Percentage of Whites with less than High School and College education, and the Percentage 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 confirm there was a statistically significant difference between clusters 0 and 2 versus clusters 1 and 3 for these population variables. On the other hand, the percentage of Hispanics, Share of 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, and Unemployment Rates.

Table 3. Differences in County-Level Variables Across Clusters

Sample Mean	Cluster 0	Cluster 1	Cluster 2	Cluster 3	*p*-value of sample Means between clusters			
					Clusters 0 and 1	Clusters 0 and 3	Clusters 1 and 2	Clusters 2 and 3
Population aged 65 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
Share of population using public transit for commuting to work (%)	0.30	1.20	0.30	3.70	<.01	<.01	<.01	<.01
Unemployment (%)	7.50	7.20	7.40	6.30	.06	<.01	.06	<.01
Voting for Clinton 2016. (%)	13.80	20.10	15.30	27.20	<.01	<.01	<.01	<.01
Less than high	13.50	11.70	12.60	11.50	<.01	.02	.01	.17

school education (%)								
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 were obtained from the Center for Systems Science and Engineering at Johns Hopkins University. 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

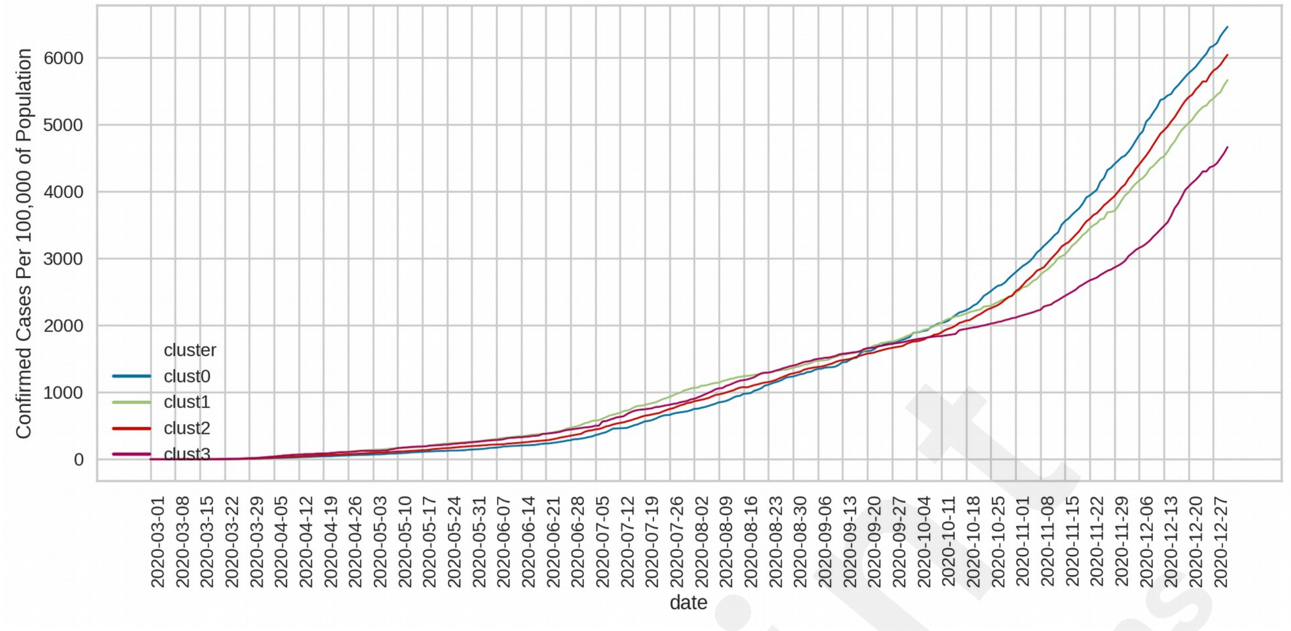
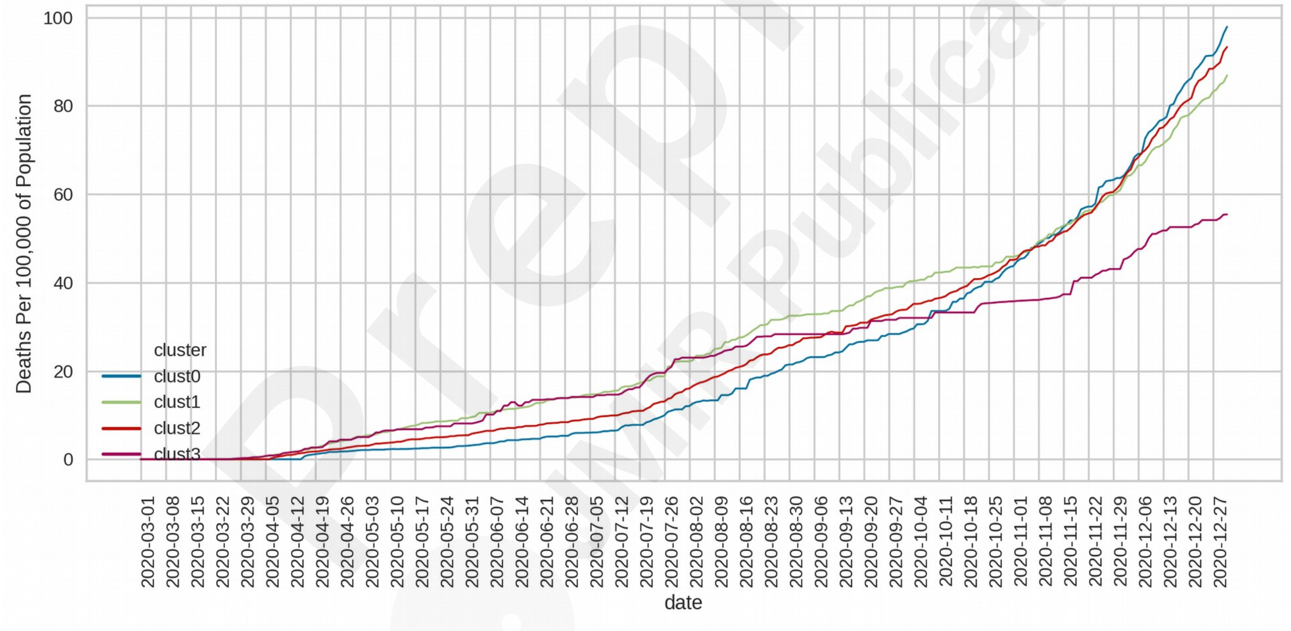


Figure 10. Daily Deaths per 100,000 Residents



DISCUSSION

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 was the first study that has used unsupervised machine learning to understand differences in population mobility across U.S. 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, out of the four 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 and over; whites with less than high school and college education; and overall population with less than a college education. Counties in these two 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 [39]. In line with this, the two 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 (2020) [42] who also use county level data during the early days of the pandemic and find 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]. While 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 [21] find that early on during the pandemic that Republican counties actually experienced lower COVID-19 cases, and therefore had lax attitudes toward mask-wearing, social distancing and lock-down 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 resulted in greater disease severity in Trump-leaning counties. In a similar vein, Allcott et al. [43] find 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 that presented here, has 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 indicates that in the US, Google Maps had 154 million users in April 2018 [41]. 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] construct a “mobility-index” and a “stay-at-home/resting-index” index based in data on almost all phone subscribers in Hungary and find 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 movements (Bryant and Elofsson[11], Askitas et al. [45]; 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 does 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 non-pharmaceutical interventions, 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 (i.e., 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 U.S. counties. This knowledge can assist policymakers in developing non-pharmaceutical interventions to restrict viral spread during the COVID-19 pandemic.

REFERENCES

1. Cucinotta D, Vanelli M. WHO Declares COVID-19 a Pandemic. Acta Biomed 2020 Mar 19;91(1):157–160. PMID:32191675

2. Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19) [Internet]. [cited 2021 Jan 4]. Available from: https://www.who.int/publications/i/item/report-of-the-who-china-joint-mission-on-coronavirus-disease-2019-(covid-19)

3. CDC. Social Distancing [Internet]. 2020 [cited 2021 Jan 4]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/social-distancing.html

4. Emergency Preparedness and Response [Internet]. 2021 [cited 2021 Jul 10]. Available from: https://emergency.cdc.gov/

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 Wiley; 2020 Aug 3;59(1):29–52. PMID:32836519

6. Rosenblatt K. A summer of digital protest: How 2020 became the summer of activism both online and offline [Internet]. NBC News. 2020 [cited 2021 Jul 10]. Available from: https://www.nbcnews.com/news/us-news/summer-digital-protest-how-2020-became-summer-activism-both-online-n1241001

7. The New York Times. See Reopening Plans and Mask Mandates for All 50 States. The New York Times [Internet] The New York Times; 2021 Jul 1 [cited 2021 Jul 10]; Available from: https://www.nytimes.com/interactive/2020/us/states-reopen-map-coronavirus.html

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 3;117(44):27087–27089. PMID: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 [Internet] 2020 Jun 15;15(1). PMID:32575957

10. Li Z, Li X, Porter D, Zhang J, Jiang Y, Olatosi B, Weissman S. 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 JMIR Research Protocols; 2020 Dec 18;9(12):e24432.

11. Bryant P, Elofsson A. Estimating the impact of mobility patterns on COVID-19 infection rates in 11 European countries. PeerJ PeerJ Inc.; 2020 Sep 15;8:e9879.

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. PMID: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. PMID: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. PMID: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. PMID: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. p. e111. PMID:32946835

17. Karaivanov A, Lu SE, Shigeoka H, Chen C, Pamplona S. Face masks, public policies and slowing the spread of COVID-19: Evidence from Canada [Internet]. bioRxiv. medRxiv; 2020. [doi: 10.1101/2020.09.24.20201178]

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 [Internet] 2020 Oct 29;17(21). PMID:33137958

19. Glaeser EL, Gorback C, Redding SJ. 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.

20. Gao S, Rao J, Kang Y, Liang Y, Kruse J, Dopfer D, Sethi AK, Mandujano Reyes JF, Yandell BS, Patz JA. 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 1;3(9):e2020485. PMID:32897373

21. Desmet K, Wacziarg R. Understanding Spatial Variation in COVID-19 across the United States. J Urban Econ 2021 Mar 11;103332. PMID:33723466

22. 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 Sep 22;15(9):e0239572. PMID:32960932

23. 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. PMID:32934147

24. Knittel CR, Ozaltun B. What Does and Does Not Correlate with COVID-19 Death Rates [Internet]. National Bureau of Economic Research; 2020 Jun. Report No.: w27391. [doi: 10.3386/w27391]

25. McLaren J. Racial Disparity in COVID-19 Deaths: Seeking Economic Roots with Census data [Internet]. National Bureau of Economic Research; 2020 Jun. Report No.: w27407. [doi: 10.3386/w27407]

26. Brown CS, Ravallion M. Inequality and the Coronavirus: Socioeconomic Covariates of Behavioral Responses and Viral Outcomes Across US Counties [Internet]. National Bureau of Economic Research; 2020 Jul. Report No.: w27549. [doi: 10.3386/w27549]

27. COVID-19 Map - Johns Hopkins Coronavirus Resource Center [Internet]. [cited 2021 Jul 10].

Available from: https://coronavirus.jhu.edu/map.html

28. 2018-elections-unoffical [Internet]. Github; [cited 2021 Jan 4]. Available from: https://github.com/MEDSL/2018-elections-unoffical

29. Song C, Liu F, Huang Y, Wang L, Tan T. Auto-encoder Based Data Clustering. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications Springer Berlin Heidelberg; 2013. p. 117–124.

30. 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 Nov 20;6:115. PMID:33501130

31. Jalali N, Sahu KS, Oetomo A, Morita PP. 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 and uHealth JMIR mHealth and uHealth; 2020 Nov 13;8(11):e21209.

32. Chung NC, Mirza B, Choi H, Wang J, Wang D, Ping P, Wang W. Unsupervised classification of multi-omics data during cardiac remodeling using deep learning. Methods 2019 Aug 15;166:66–73. PMID:30853547

33. Awesome P. Variational Recurrent Autoencoder for timeseries clustering in pytorch [Internet]. Python Awesome; 2019 [cited 2021 Jul 10]. Available from: https://pythonawesome.com/variational-recurrent-autoencoder-for-timeseries-clustering-in-pytorch/

34. Yu X, Li H, Zhang Z, Gan C. The Optimally Designed Variational Autoencoder Networks for Clustering and Recovery of Incomplete Multimedia Data. Sensors [Internet] 2019 Feb 16;19(4). PMID:30781499

35. Keras Team. Keras: the Python deep learning API [Internet]. [cited 2020 May 23]. Available from: https://keras.io/

36. Using the elbow method to determine the optimal number of clusters for k-means clustering [Internet]. [cited 2021 Jul 10]. Available from: https://bl.ocks.org/rpgove/0060ff3b656618e9136b

37. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987 Nov 1;20:53–65.

38. Identifying Feature Relevance using a Random Forest [Internet]. [cited 2021 Jul 10]. Available from: http://videolectures.net/slsfs05_rogers_ifrur/

39. Jones B. An examination of the 2016 electorate, based on validated voters [Internet]. 2018 [cited 2021 Jul 10]. Available from: https://www.pewresearch.org/politics/2018/08/09/an-examination-of-the-2016-electorate-based-on-validated-voters/

40. Sanchez B. Trump supporter on not wearing a mask: It’s a fake pandemic. CNN [Internet] 2020 Sep 11 [cited 2021 Jul 10]; Available from: https://www.cnn.com/videos/politics/2020/09/11/trump-rally-attendees-michigan-ctn-vpx.cnn

41. Most popular social media apps in U.S [Internet]. [cited 2021 Jul 10]. Available from: https://www.statista.com/statistics/248074/most-popular-us-social-networking-apps-ranked-by-audience/

42. Zidian X, Dongmei L. Health and Demographic Impact on COVID-19 Infection and Mortality in US Counties. medRxiv 2020.05.06.20093195; doi: https://doi.org/10.1101/2020.05.06.20093195.

43. 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. doi: 10.1016/j.jpubeco.2020.104254. Epub 2020 Aug 6. PMID: 32836504; PMCID: PMC7409721.

44. Szocska, M., Pollner, P., Schiszler, I. et al. Countrywide population movement monitoring using mobile devices generated (big) data during the COVID-19 crisis. Sci Rep 11, 5943 (2021). https://doi.org/10.1038/s41598-021-81873-6.

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 11, 1972 (2021). https://doi.org/10.1038/s41598-021-81442-x.

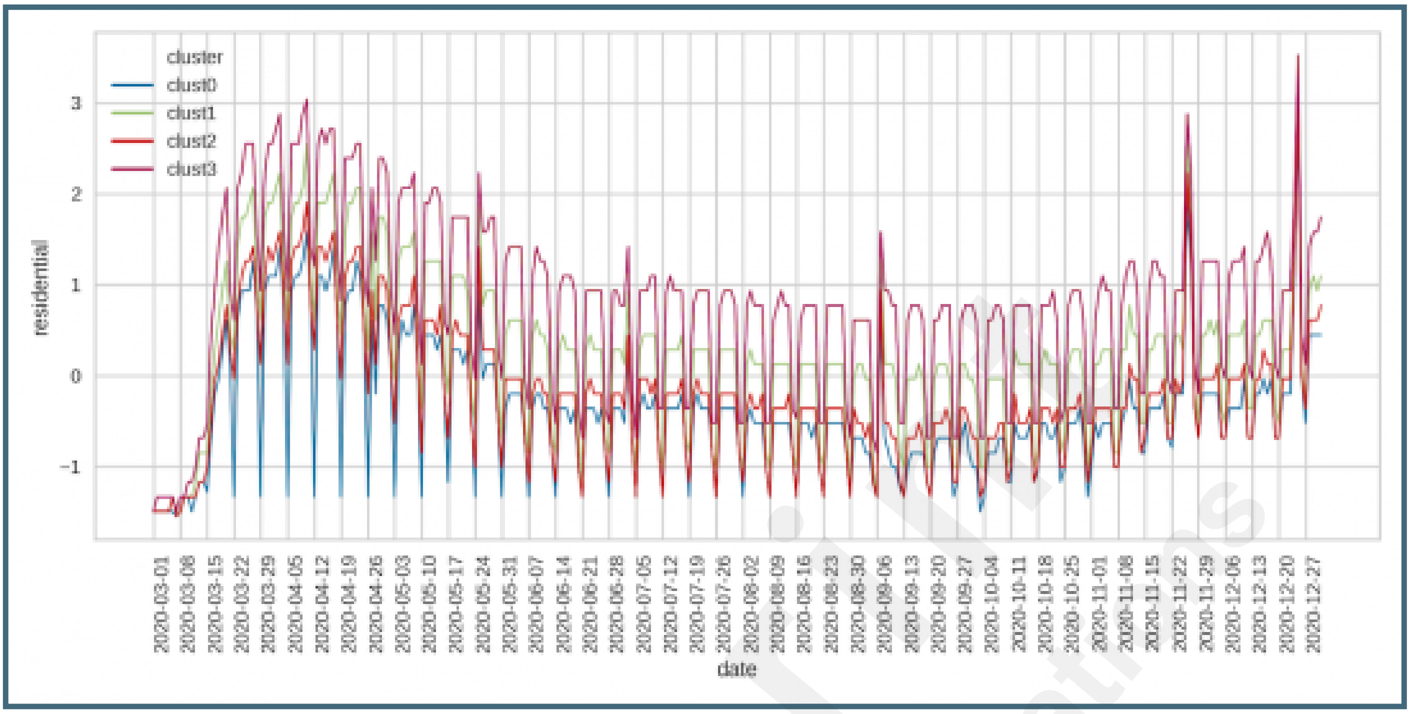
46. Stevens NT, Sen A, Kiwon Francis. Morita PP., Steiner SH, Zhang Q> Estimating the effects of Non-Pharmaceutical Interventions (NPIs) and population mobility on daily COVID-19 cases: evidence from Ontario}. Canadian Public Policy (2021). https://doi.org/10.3138/cpp.2021-022.

Supplementary Files

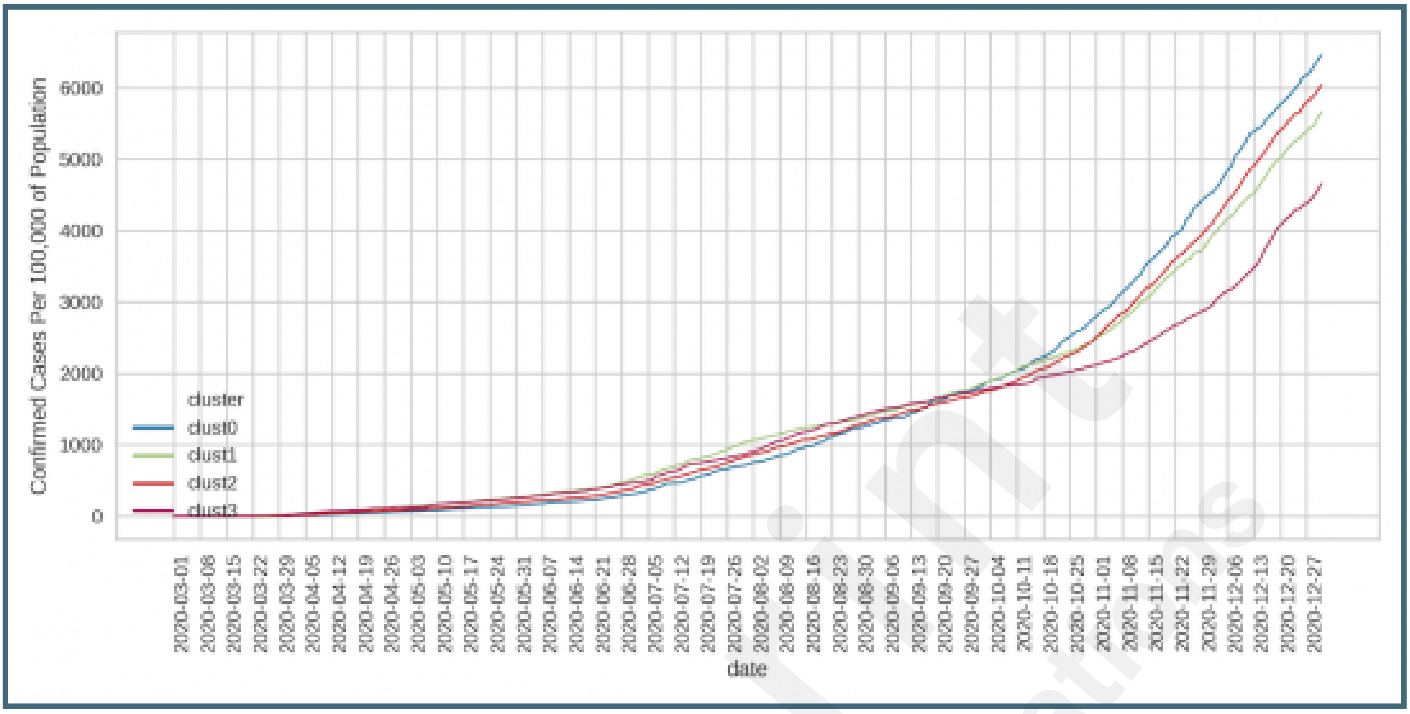
Untitled.
URL: http://asset.jmir.pub/assets/06697812d5b64a6548c8f67f8e16031e.docx

Figures

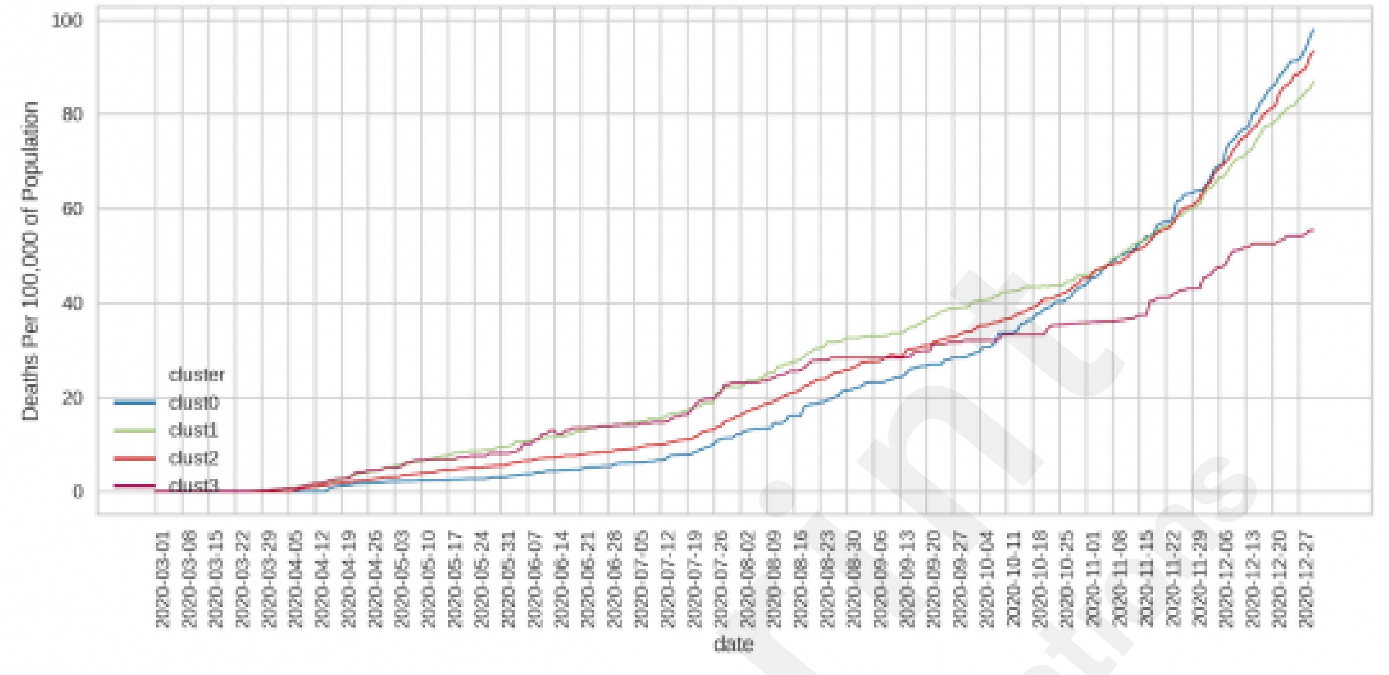
Untitled.


cluster
clust0
clust1
clust2
clust3
residential
3
2
1
0
−1
date
2020-03-01
2020-03-08
2020-03-15
2020-03-22
2020-03-29
2020-04-05
2020-04-12
2020-04-19
2020-04-26
2020-05-03
2020-05-10
2020-05-17
2020-05-24
2020-05-31
2020-06-07
2020-06-14
2020-06-21
2020-06-28
2020-07-05
2020-07-12
2020-07-19
2020-07-26
2020-08-02
2020-08-09
2020-08-16
2020-08-23
2020-08-30
2020-09-06
2020-09-13
2020-09-20
2020-09-27
2020-10-04
2020-10-11
2020-10-18
2020-10-25
2020-11-01
2020-11-08
2020-11-15
2020-11-22
2020-11-29
2020-12-06
2020-12-13
2020-12-20
2020-12-27

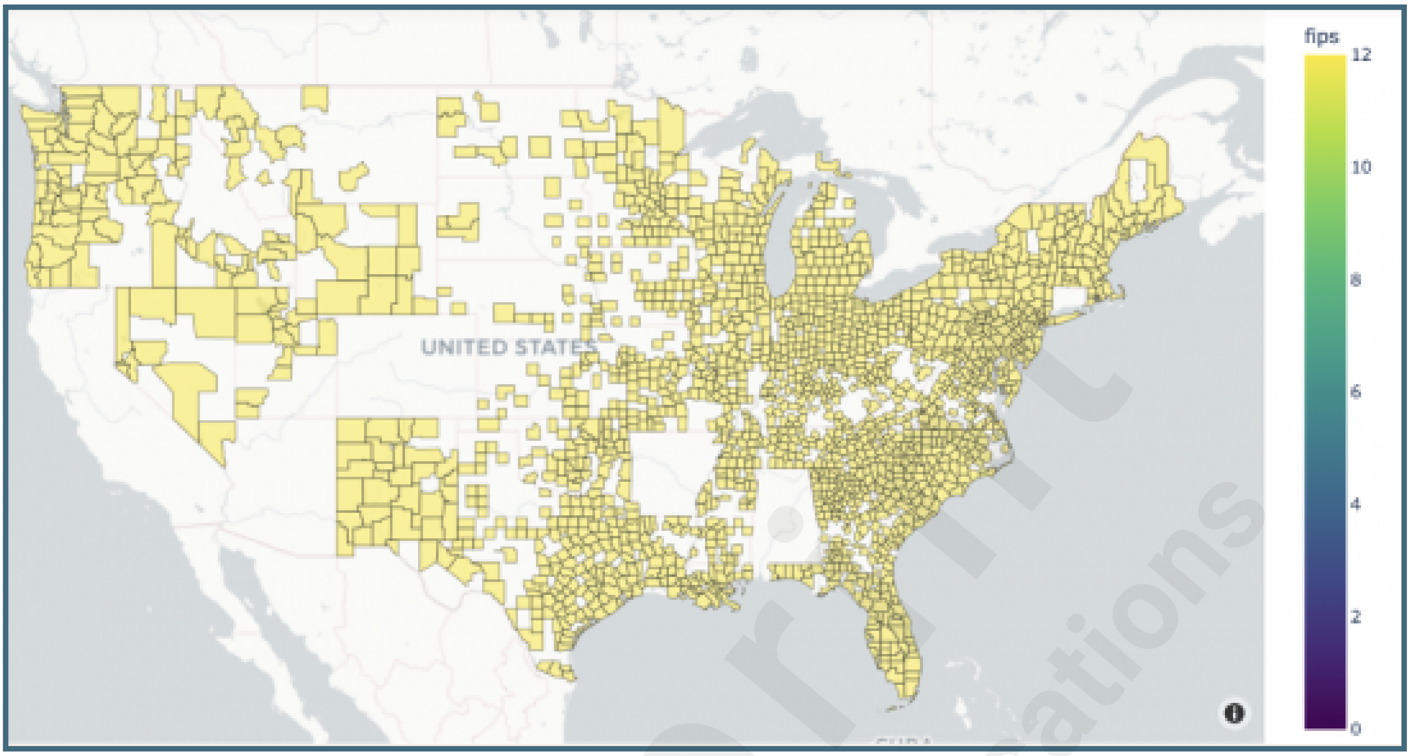
Untitled.


Confirmed Cases Per 100,000 of Population
0
1000
2000
3000
4000
5000
6000
cluster
clust0
clust1
clust2
clust3
2020-03-01
2020-03-08
2020-03-15
2020-03-22
2020-03-29
2020-04-05
2020-04-12
2020-04-19
2020-04-26
2020-05-03
2020-05-10
2020-05-17
2020-05-24
2020-05-31
2020-06-07
2020-06-14
2020-06-21
2020-06-28
2020-07-05
2020-07-12
2020-07-19
2020-07-26
2020-08-02
2020-08-09
2020-08-16
2020-08-23
2020-08-30
2020-09-06
2020-09-13
2020-09-20
2020-09-27
2020-10-04
2020-10-11
2020-10-18
2020-10-25
2020-11-01
2020-11-08
2020-11-15
2020-11-22
2020-11-29
2020-12-06
2020-12-13
2020-12-20
2020-12-27
date

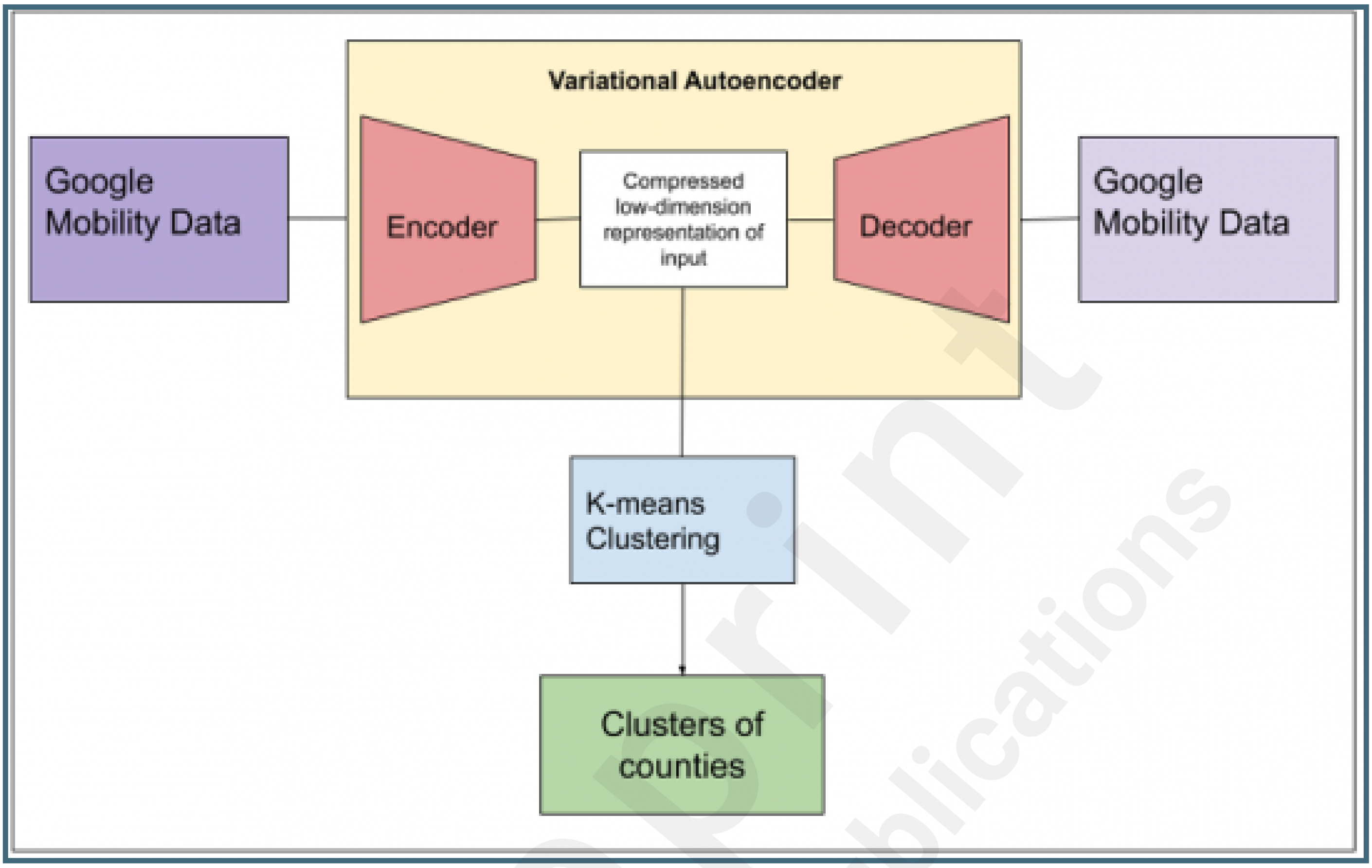
Untitled.


100
80
60
40
20
0
Deaths Per 100,000 of Population
cluster
clust0
clust1
clust2
clust3
2020-03-01
2020-03-08
2020-03-15
2020-03-22
2020-03-29
2020-04-05
2020-04-12
2020-04-19
2020-04-26
2020-05-03
2020-05-10
2020-05-17
2020-05-24
2020-05-31
2020-06-07
2020-06-14
2020-06-21
2020-06-28
2020-07-05
2020-07-12
2020-07-19
2020-07-26
2020-08-02
2020-08-09
2020-08-16
2020-08-23
2020-08-30
2020-09-06
2020-09-13
2020-09-20
2020-09-27
2020-10-04
2020-10-11
2020-10-18
2020-10-25
2020-11-01
2020-11-08
2020-11-15
2020-11-22
2020-11-29
2020-12-06
2020-12-13
2020-12-20
2020-12-27
date

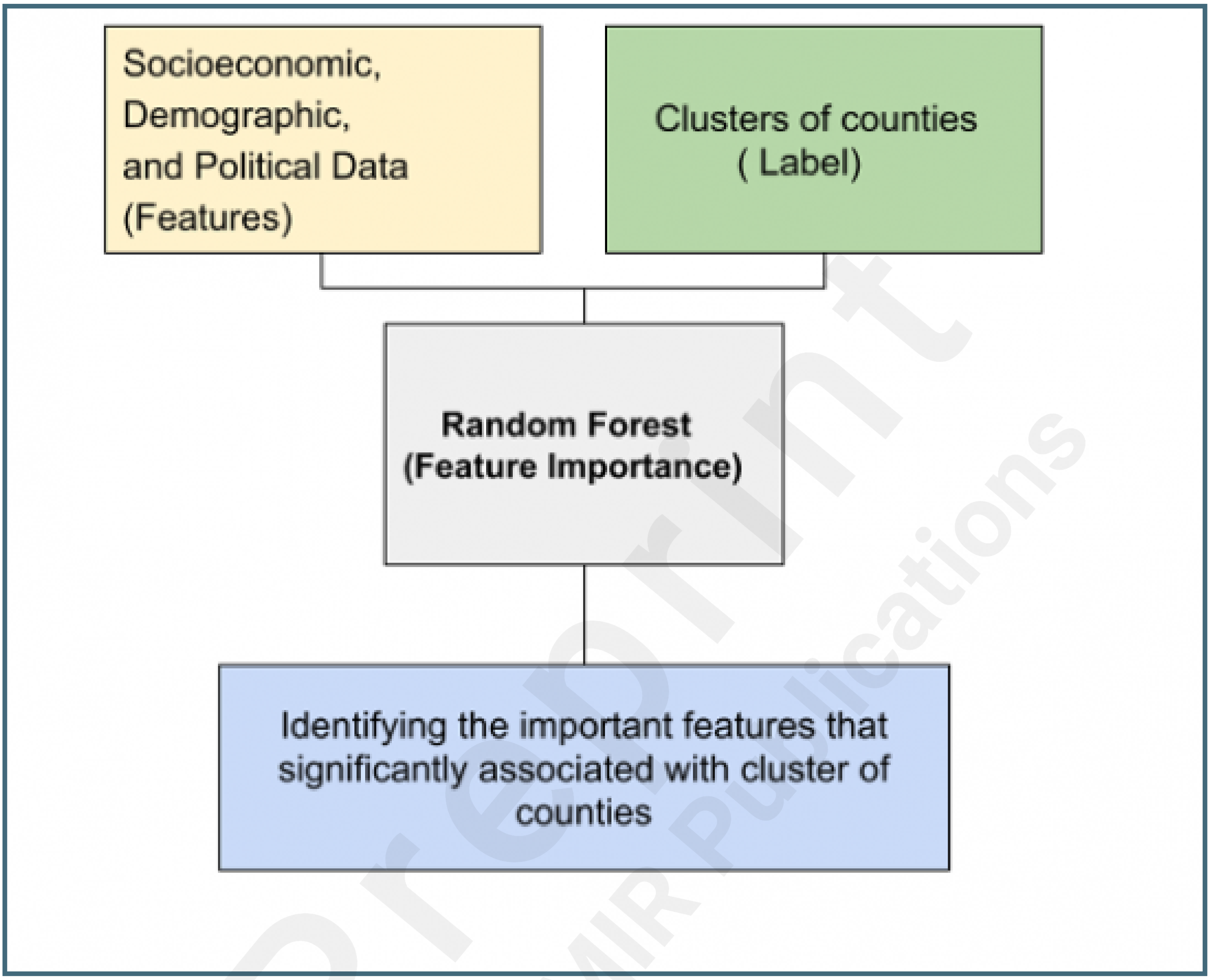
Distortion Score Elbow for K-means Clustering.



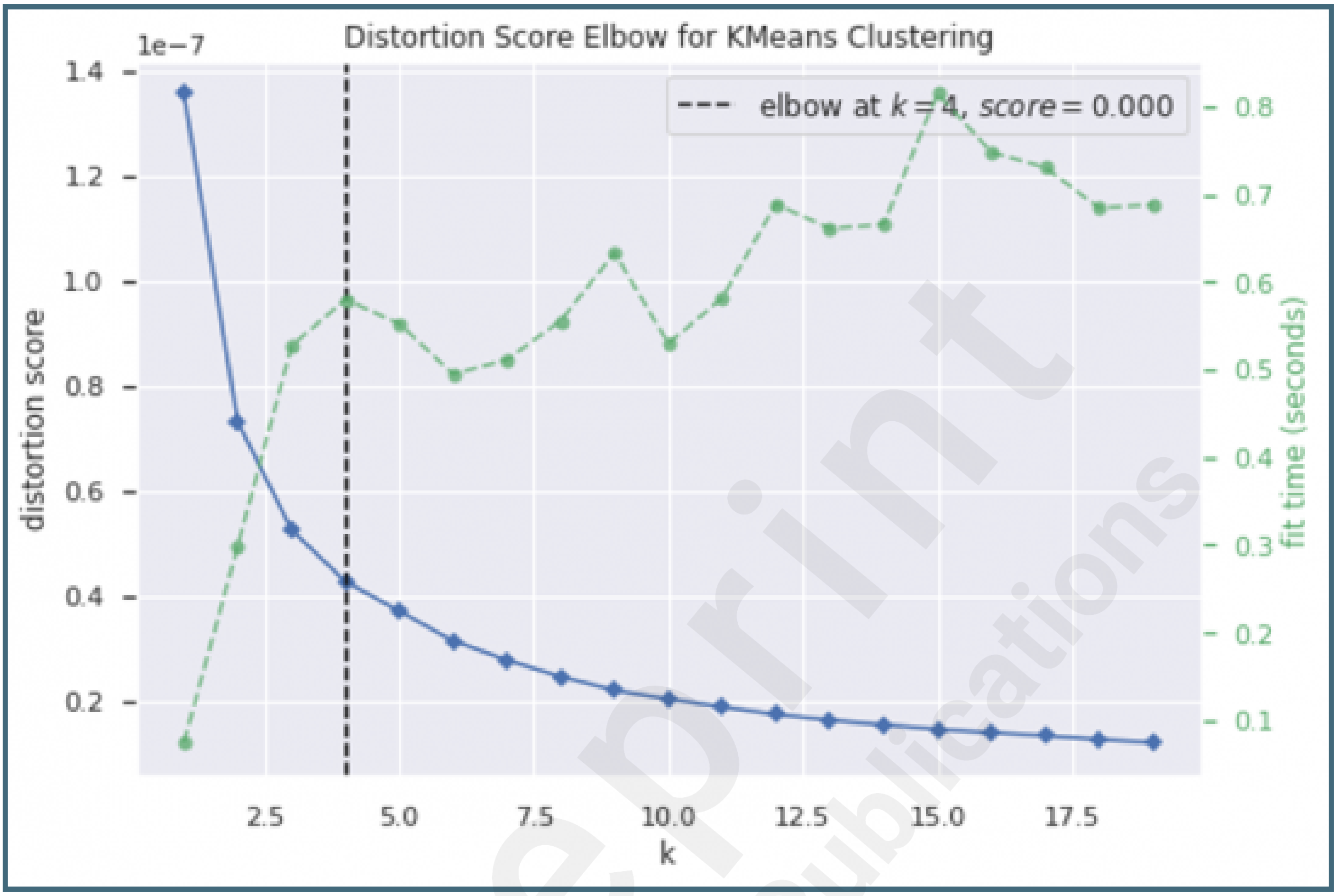
Google Retail Mobility.



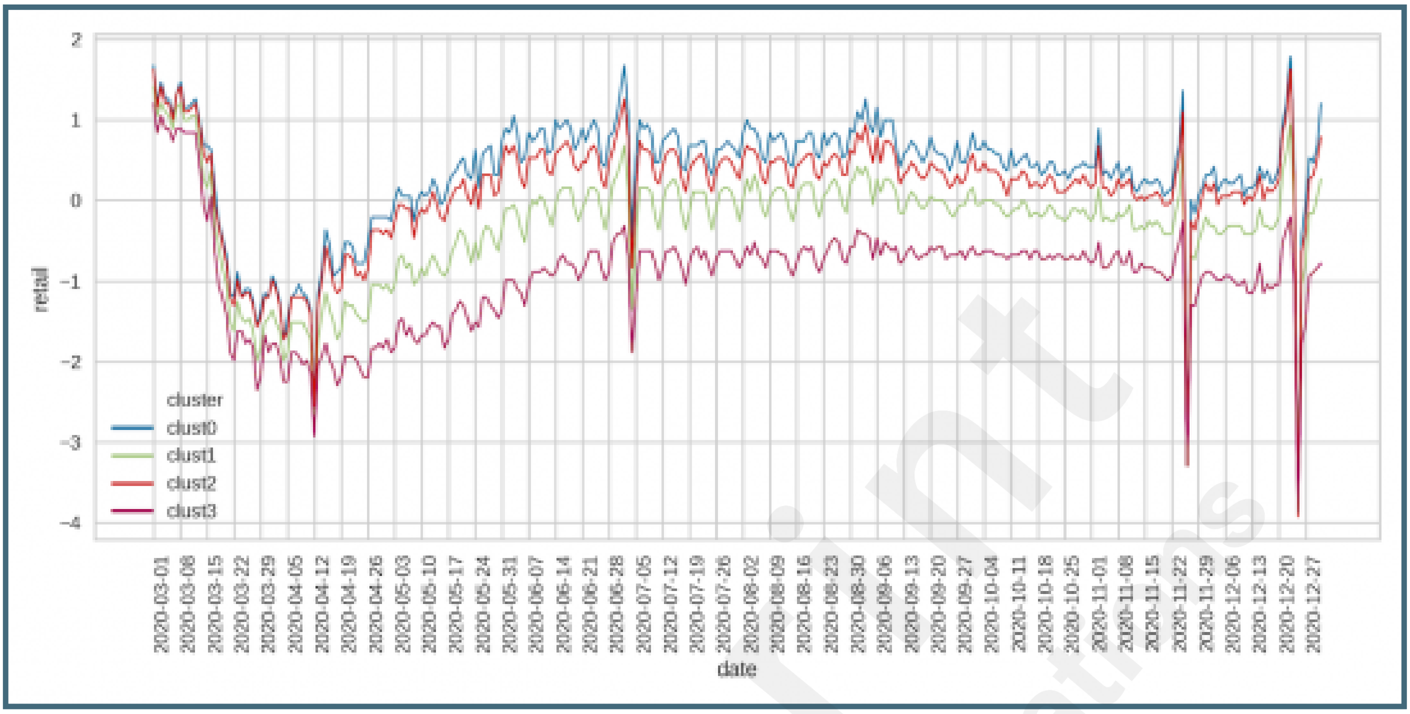
Google Groceries & Pharmacies Mobility.


Socioeconomic, Demographic, and Political Data (Features)
Clusters of counties (Label)
Random Forest (Feature Importance)
Identifying the important features that significantly associated with cluster of counties

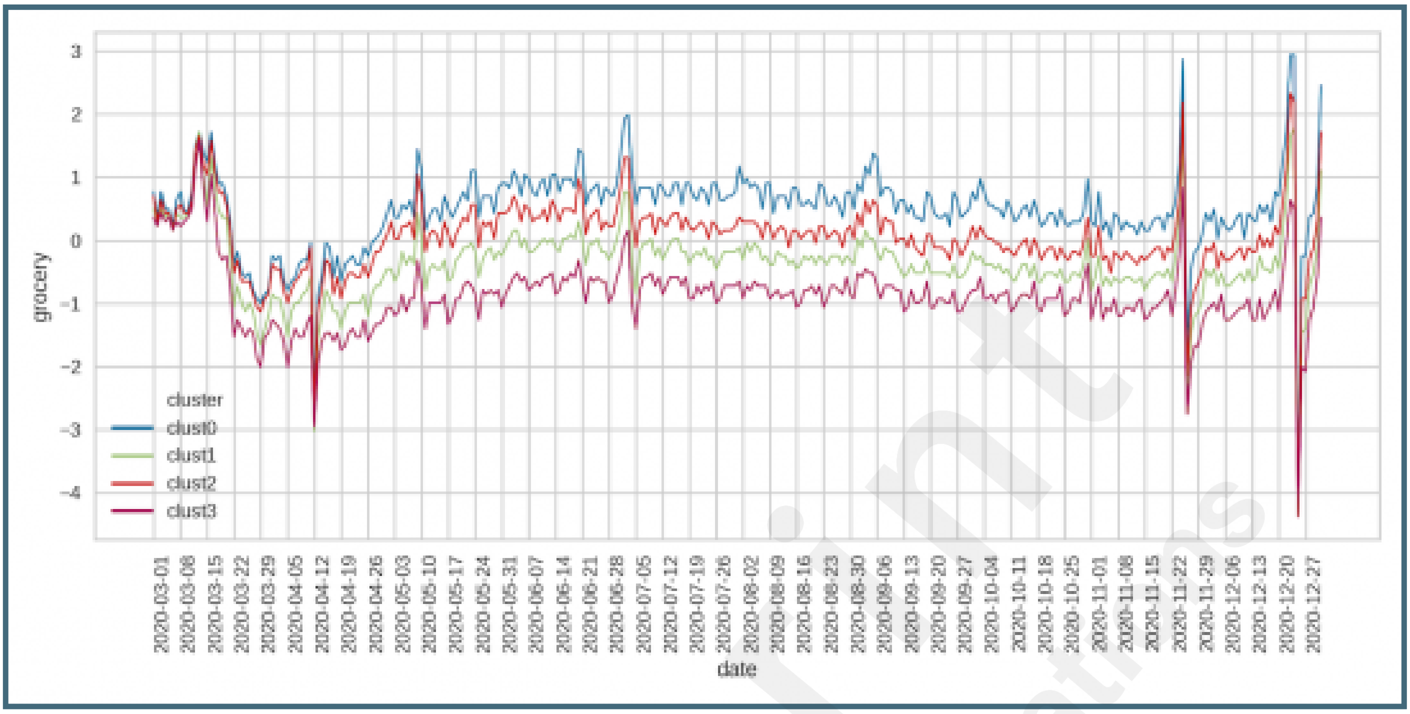
Google Workplace Mobility.



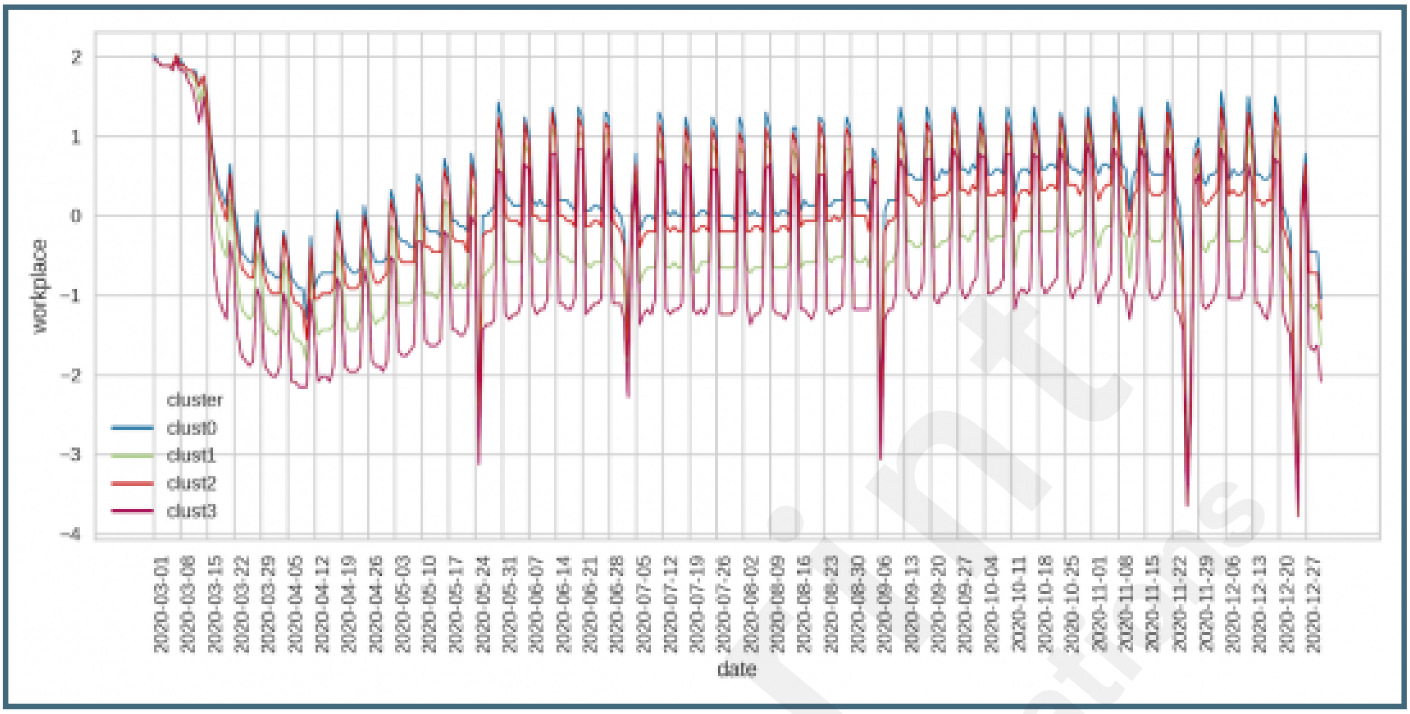
Residential Mobility.



Daily Cases per 100,000 Residents.



Daily Deaths per 100,000 Residents.


workplace
cluster
clust0
clust1
clust2
clust3
date