Socio-economic impact on COVID-19 cases and deaths and its evolution in New Jersey

Dhammika Amaratunga[1] · Javier Cabrera[2] · Debopriya Ghosh[4] ·
Michael N. Katehakis[2] · Jin Wang[3] · Wenting Wang[2]

Abstract

Socio-economic factors could impact how epidemics spread. In this study, we investigated the possible effect of several local socio-economic factors on the case count and time course of confirmed Covid-19 cases and Covid-19-related deaths across the twenty one counties of New Jersey. Socio-economic and geographic factors considered included population, percentage of elders in the population, percentage of low-income households, access to food and health facilities and distance to New York. We found that the counties could be clustered into three groups based on (a) the case totals, (b) the total number of deaths, (c) the time course of the cases and (d) the time course of the deaths. The four groupings were very similar to one another and could all be largely explained by the county population, the percentage of low-income population, and the distance of the county from New York. As for food and health factors, the numbers of local restaurants and pharmacies significantly influenced the total number of cases and deaths as well as the epidemic's evolution. In particular, the number of cases and deaths showed a decrease with the number of McDonald's within the county in contrast to other fast-food or non-fast food restaurants. Overall, our study found that the evolution of the epidemic was influenced by certain socio-economic factors, which could be helpful for the formulation of public health policies.

Keywords COVID-19 · Time course data · Cluster analysis · Simpson's rule dissimilarity · Multidimensional scaling · Linear model

1 Introduction

Socio-economic factors have been found to impact how epidemics spread. The availability of Covid-19 related data at a small geographic scale has enabled us to compare and contrast disease evolution across different geographic locations and study its relation to local socioeconomic factors.

Socio-economic factors measure different components and aspects of the social environment, which often include occupation, income, education, facilities, etc. cf. Krieger (2001).

Emerging infectious diseases have been found to be greatly driven by socio-economic, environmental and eco-logical factors (Jones et al. 2008; Khalatbari-Soltani et al. 2020). Khalatbari-Soltani et al. (2020) recommended that it was crucial to collect data on socio-economic determinants in COVID-19 studies. Some studies have examined socio-economic impacts on COVID-19, e.g., income (Kim and Bostwick 2020), education (Kim and Bostwick 2020; Drefahl et al. 2020), living condition (Raisi-Estabragh et al. 2020), population density (Gu et al. 2020). Raisi-Estabragh et al. (2020) found that the factors underlying ethnic differences in COVID-19 are socio-economic status, e.g., living conditions. McLaren (2020) identified the socioeconomic roots of racial disparities in COVID-19 mortality. It was found that the ethnic differences in COVID-19 disappeared when factors related to education, occupation, and commuting patterns were controlled. Kim and Bostwick (2020) created a social vulnerability index (SVI) by using data about poverty, education, female-headed households with children, median household income, and employment ratio. They found that SVI is associated with COVID-19 deaths in Chicago. Cordes and Castro (2020) conducted a similar spatial analysis for New York City. Gu et al. (2020) identified that high population density was relevant to COVID-19 outcomes in Michigan. Additionally, geographic proximity to food (fast food restaurants, etc) and to health facilities (e.g., fitness centers) have been studied extensively as part of research on environmental health and food security among others (Aztsop and Joy 2013; Hilmers et al. 2012). The importance of access to food in Chicago (Kim and Bostwick 2020), New York (Cordes and Castro 2020) and Michigan (Gu et al. 2020) suggests that it should also be relevant to the development of COVID-19. This same premise can be applied more to access to health facilities such as pharmacies, fitness centers, etc.

Given the importance of understanding how particular socioeconomic factors could affect the spread of COVID-19, in this paper, we focus on the evolution of COVID-19 confirmed cases and deaths in the counties of the state of New Jersey (NJ). The socio-economic factors considered in this study include: (a) demographic variables (i.e., population, percentage of population older than 80, percentage of low-income population), (b) a geographic variable (i.e., distance from each county to New York), and (c) variables about geographic access to food and health facilities (i.e., the number of fast foods, restaurants, nursing homes, groceries, fitness centers, and pharmacies in the counties). We first examined the impacts of these socio-economic factors on the total number of cases and deaths, then the impacts on the dissimilarity between the time profiles of the counties.

The rest of the paper is organized as follows. In Sect. 2 we discuss the sources of data and the statistical methodology that was employed. In Sect. 3 we present the results of the statistical analysis, followed by the discussion and conclusions in Sects. 3 and 5 respectively.

2 Methods

2.1 Data source

We collected the daily cumulative counts of confirmed COVID-19 cases and deaths in each county of NJ during the period March 1st to September 9th, 2020. The time-course data was collected from the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) (Dong et al. 2020). The population in each county of NJ was also available. The percentage of population older than 80 was obtained from the 2019 population esti-

mates for NJ counties (https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/counties/asrh/cc-est2019-agesex-34.csv).

Data for the annual total household income data was collected from the United States Census Bureau website (data.census.gov) by the American Community Survey (https://data.census.gov/cedsci/table?t=IncomeandPoverty&g=0400000US34,34.050000&tid=ACSDP1Y2018.DP03&moe=false&hidePreview=true). Then the estimated number and proportion of households within a certain income range for a specific county can be readily computed. Families with household incomes less than 80% of the local median income are designated as low-income families. Then, the proportion of local low-income households for each county can be obtained.

As for distance from each county to New York City (NYC), we first obtained coordinates of the center of each county and NYC by the Google Map api. The distance was obtained by using the *geopy* site (https://geopy.readthedocs.io/en/stable/).

The food and health facilities considered in the study included McDonald's, other fast-food, non-fast-food restaurants, grocery, fitness, pharmacy. Since McDonald's is an icon of fast-food, we separated it from other fast-food. We collected the data about geographic access to the facilities by querying the Yelp Fusion API (https://www.yelp.com/developers/documentation/v3). Facilities could be searched based on the provided coordinate and radius. We divided NJ into 4000×4000 meters grids, searched facilities in each grid by providing the coordinate of the center and the radius of the circumscribed circle of the grid, and then removed duplicate results. With the address and zip code in the search result, which county the facilities were located was determined. At last, we obtained the number of the facilities in each county.

To find the list of all nursing homes we searched the state of New Jersey department of health website (https://healthapps.state.nj.us/Facilities/fsSearch.aspx). We also included residential health care and assisted living facilities.

2.2 Explanatory variables

The explanatory variables included in this study are food-and-health access variables (i.e., the number of McDonald's x_1, other fast food x_2, non-fast-food restaurants x_3, groceries x_4, fitness centers x_5, pharmacies x_6, and nursing homes x_7), demographic variables (i.e., population, p; the percentage of population older than eighty, o, referred as *elder* for short; and the proportion of low-income households in each county, l, referred as *low-income* for short), the geographic variable (i.e., distance to NYC), d.

Some variables were highly correlated (see Fig. 4 in "Appendix A", the matrix of scatter plots for the data). Particularly, the food-and-health access variables, $x_1, \ldots, x_7$, are greatly related to population (p). To reduce this effect for analysis, two new datasets were created. In the first one, the food-and-health access variable was divided by population. In the second one, we ran linear regression for these variables with population and distance to NYC since the food-and-health variables were also greatly related to distance to NYC (d), then substituted the variables with the corresponding residuals. The matrices of scatter plots of the two datasets were shown in "Appendix A", in which the variables were comparatively independent.
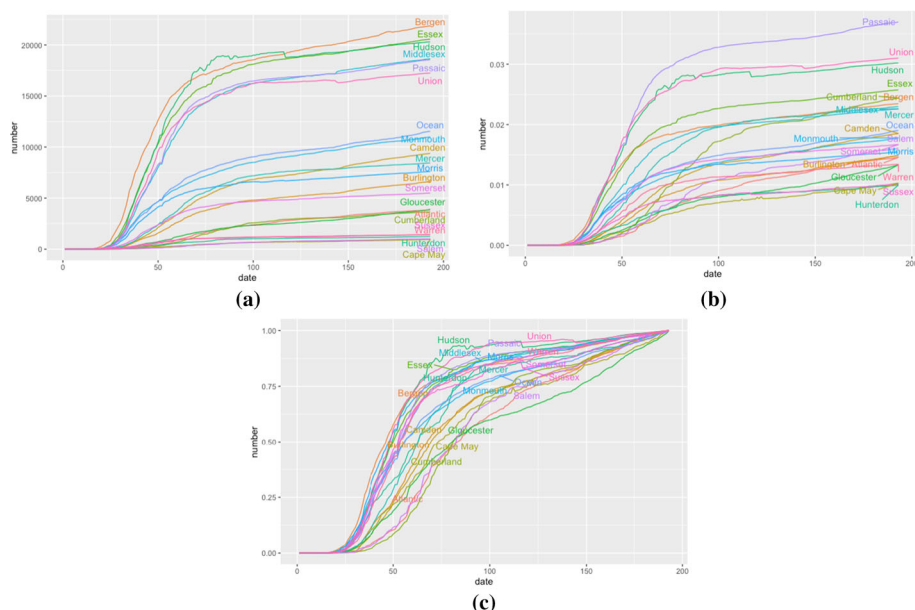
Fig. 1 The different time-course data relevant to cases. **a** The raw time-course data, $\{c_{ij}\}$; **b** the time-course data divided by population, c_{ij}/p_j; **c** the time-course data divided by total number of cases, c_{ij}/c_{mj}

2.3 Statistical analysis

In this study, we analyzed both the total number of confirmed cases and the time-course data. The time-course data was of the form (t_i, c_{ij}) where t_i was the ith date $(i = 1, \ldots, m)$ and c_{ij} was the total number of cases up to date t_i in County j.

2.3.1 Analysis on total number of cases

Since the number of cases would clearly be related to population, we also used the proportion of cases within a county as response variable, i.e. $y_j = c_{mj}/p_j$, where p_j is the population of County j. We ran a linear regression model and then backward stepwise regression to select variables.

The total number of cases was highly related to population and distance to NYC (see Fig. 4 in "Appendix A"). We investigated also how these socio-economic factors could be contributing to total cases after removing population and distance to NYC as factors. This was done by modeling the residuals from fitting $\{c_{mj}\}$ vs $\{p_j\}$ and $\{d_j\}$.

2.3.2 Analysis of the time-course data

The time-course data was shown in Fig. 1a. It was also analyzed as a proportion of the population, which was obtained by letting $y_{ij} = c_{ij}/p_j$. The profiles were shown in Fig. 1b. The cumulative cases were also analyzed as a proportion of the total number of cases, that is, $y_{ij} = c_{ij}/c_{mj}$ to measure the gradients of the time profiles. The profiles were shown in Fig. 1c respectively. We performed the following analysis for the profiles.
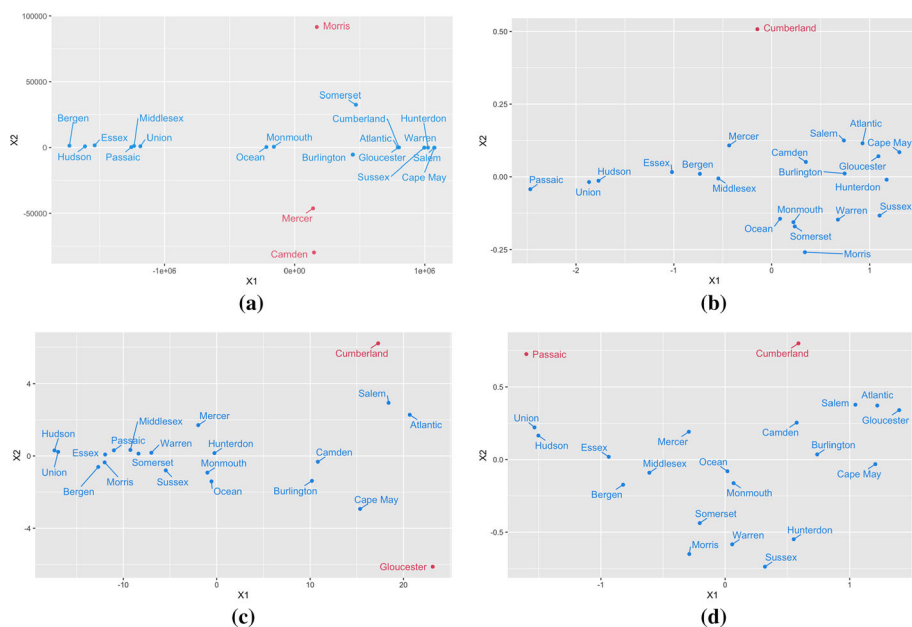
(a) Dissimilarity matrix

Fig. 2 Scatter plots of the first two components of multidimensional scaling for dissimilarity matrix relevant to **a** the raw time-course case data, $\{c_{ij}\}$; **b** the time-course case data divided by population, $\{c_{ij}/p_j\}$; **c** the time-course case data divided by total number of cases, $\{c_{ij}/c_{mj}\}$; **d** combination of $\{c_{ij}/p_j\}$ and $\{c_{ij}/c_{mj}\}$

The dissimilarity between the time profiles for counties j and k can be assessed by the area between the jth and kth curves. A simple way to define the area between the curves is by the formula:

$$D_{jk} = \sum_i |y_{ij} - y_{ik}|\delta_i \qquad (1)$$

where $\delta_i = t_i - t_{i-1}$, and since in our case t_i represents the day then $\delta_i = 1$. Another similarity measure based on area between the curves is derived from Simpson's rule:

Let $z_{ijk} = |y_{ij} - y_{ik}|$ and let $\delta_i = \delta$

$$D_{jk} = \frac{\delta}{3}(z_{1jk} + 4z_{2jk} + 2z_{3jk} + 4z_{4jk} + 2z_{5jk} + \cdots + 4z_{(n-1)jk} + z_{njk}) \qquad (2)$$

This later definition we called Simpson's rule similarity gives a better approximation to the area between the curves. However, because our times series is quite long, the difference between the two formulas is small.

(b) Multidimensional scaling

Multidimensional scaling (MDS) can then be applied to find a set of points $\{a_j\}$ such that the Euclidean distance between a_j and a_k approximates D_{jk}. Figure 2a–c showed a scatterplot of the points $\{a_j\}$ for the different time profiles. Some outliers were marked in red.

In addition, the two sets of dissimilarities, denoted $D_{jk}(P)$ and $D_{jk}(T)$, can be combined to give an idea of the overall picture of the differences between counties that captured both

the cumulative counts as well as the time profiles:

$$D_{jk}(O) = \frac{D_{jk}(P)}{\sqrt{var(D_{jk}(P))}} + \frac{D_{jk}(T)}{\sqrt{var(D_{jk}(T))}} \tag{3}$$

Multidimensional scaling (MDS) of $\{D_{jk}(O)\}$ generated the set of points for the counties in Fig. 2d. Two outlier counties, Passaic and Cumberland, were marked in red.

(c) The first component regression

Since a large proportion of the dissimilarity between counties could be accounted for by the first eigenvector of MDS (larger than 0.90), it was reasonable to regard the values $\{a_j\}$ along this eigenvector as carrying the most information regarding differences between counties; these values were denoted $\{a_j^*\}$, which were then used as response variable and modeled against the explanatory variables. We also investigated how socio-economic factors could be contributing to differences in spread of Covid-19 after removing population and distance to NYC as factors. This was done by modeling the residuals from fitting $\{a_j^*\}$ vs $\{p_j\}$ and $\{d_j\}$.

An identical analysis was done with the cumulative deaths data. The corresponding plots could be found in "Appendix B".

3 Results

Table 1 showed the linear regression results using the data set where the food-and-health variables (i.e., $x_1, x_2, \ldots, x_7$) were divided by population (p). Columns 2-5 demonstrated the regression results for the total number of cases and deaths (Column C and D), as well as their proportion (Column C/P and D/P). The effects were almost identical. The most important factors were McDonalds, other fast-food, fitness, pharmacy, low-income and population. As expected, the total numbers (of cases and deaths) significantly increased with low-income and population density. For food factors, other fast food restaurants were positive with p-value < 0.001 for cases and p-value < 0.01 for deaths, while McDonalds was negative (with p-value < 0.01). This indicates that McDonalds in contrast to other fast-food and non-fast-food outlets was associated with lower transmission of the Covid-19 virus. Surprisingly, the collected data indicated that non-fast-food restaurants and grocery stores were not significant. For health factors, fitness centers were shown to be positively correlated with p-value < 0.01 for cases and p-value < 0.05 for deaths. Further, the number of pharmacies was negatively correlated with p-value < 0.01 for cases and p-value < 0.05 for deaths. Surprisingly, our data indicated that the number of nursing homes was not significant. As for the analysis on the proportion of cases, the effects of the factors were almost the same as that on the total cases (Column C). However, the regression results on the proportion of deaths were much different from the others. In particular, it was dominated by distance to NYC, while food-and-health factors were not significant except the number of nursing homes.

The other columns in Table 1 demonstrated the regression analysis when using the first component of MDS for the dissimilarity matrix. This analysis also showed that the reason for the dissimilarity on cases and deaths between counties could be explained by the socio-economic factors. The effects on cases dissimilarity (Column C) were almost identical to that on deaths (Column D). The most important factors were population, distance to NYC, and low-income. For food-and health factors, only non-fast-food was significant, which indicated that the more non-fast-food restaurants resulted in more cases and deaths. On the other hand, the results for the dissimilarity on cases (deaths) proportion (Column C/P and D/P) were

Table 1 Regression results using the data set where the food-and-health variables (i.e., $x_1, x_2, \ldots, x_7$) were divided by population (p)

Variables	Total number of				The 1st component of MDS for dissimilarity matrix about time profile of cumulative							
	C	C/P	D	D/P	C	C/P	C/T	CC	D	D/P	D/T	CD
McDonalds(x_1/p)	−0.54^	−0.14&	−0.41^			−0.74&						
Other fast-food(x_2/p)	0.38&	0.19^	0.23^			0.39$		0.21#				
Non-Fast-Food(x_3/p)			0.21	0.05	0.11#	0.48$			0.26$	0.17	0.14	0.13
Grocery(x_4/p)			−0.16				0.18#	0.18				
Fitness(x_5/p)	0.39^	0.11#	0.25$				−0.16					
Pharmacy(x_6/p)	−0.38^	−0.19$	−0.22$		−0.09	−0.41#		−0.22	−0.16			
Nursing home(x_7/p)	−0.07			0.12^						0.18		0.15
Low-income(l)	0.25#	0.21&	0.18$	0.16&	0.29&	0.41^		0.25$	0.17$	0.41&		0.23$
Elder(o)							0.19$					
Population(p)	0.52&		0.49&		0.57&		−0.		0.67&			
Distance to NYC(d)	0.35$			−0.35&	−0.37^		−1.22&	−0.85&	−0.36^	−0.97&	−0.88&	−1.01
Adjusted R-square	0.94	0.84	0.96	0.83	0.94	0.76	0.91	0.86	0.91	0.83	0.80	0.84

Notation C(D) means total number of confirmed cases (deaths); C/P (D/P) means total number of confirmed cases (deaths) divided by population; C/T (D/T) means cumulative cases (deaths) divided by total number of cases (deaths); CC (DD) means the combination for cases (deaths) analysis. Significant codes: 0 & 0.001 ∧ 0.01 $ 0.05 # 0.10
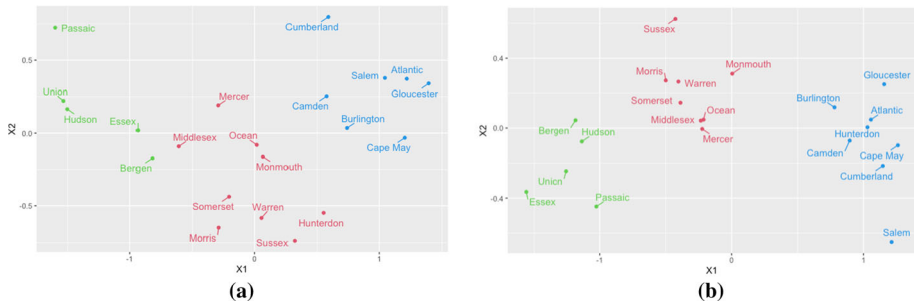
Fig. 3 K-means clustering on the first two components of the combination similarity matrix for **a** cumulative cases and **b** deaths

similar to that for total number of cases (death) (Column 3 and 5), except that fitness centers was not significant in Column C/P and nursing home was not significant for D/P. Also, it was worth noting that death proportion dissimilarity was influenced greatly by distance to NYC. The larger the distance was, the less the death proportion. Columns C/T (D/T) showed the effects on the dissimilarity on the gradients between the counties, where the gradients were computed as the ratios of the cumulative cases (deaths) divided by total number of cases (deaths) calculated in the time course of the curves. The dominant effect was the distance to NYC, as seen in Figs. 2c and 8 in "Appendix B". Furthermore, we also combined the two sets of dissimilarities (i.e., C/P and C/T) and run regressions for the first component of MDS. The results for cases and deaths were shown in Column CC and CD respectively. Again, it depended greatly on the distance to NYC, followed by low-income. For food-and health factors, only the other fast-food outlets factor was significant.

Based on the results above, the most important factors were population and distance to NYC. Hence, we removed the two factors in responses (total numbers and the first components of MDS) and the food-and-health variables. It was done by using residuals obtained by fitting them vs the two factors. After that, we conducted the same analysis. Table 2 demonstrated the results. Many socio-economic factors were still significant. In particular, more food-and-health factors came out and were significant.

We applied K-means clustering on the first two components of the combination similarity matrix $\{D_{jk}(O)\}$ to cluster the counties (see Fig. 3). The clustering results were almost identical, except for the county of Hunterdon.

4 Discussion

We may expect that the places where people got food, e.g., restaurant or grocery, were the main locations the virus spread. Grocery should be better than restaurant since grocery was less crowded. In our analysis, grocery was not significant. However, one interesting finding in this study was that the effects of different restaurant types were different. In most models, we found a contrast between the frequency of McDonald's vs other fast-food or non-fast-food restaurants. The frequency of McDonald's was associated with fewer cases (deaths), while other fast-food or non-fast-food restaurants was associated with more cases (deaths). One explanation might be that McDonald's did a good job on hygiene or crowd control measures.

The effects of most factors were consistent across all the analysis. Pharmacies were consistently associated with fewer cases and deaths, whereas gyms and percent of low-income

Table 2 Regression results using the dataset where variables $x'_1, x'_2, \ldots, x'_7$ were the residuals obtained from fitting food-and-health variables $x_1, x_2, \ldots, x_7$ versus population (p) and distance to NYC (d)

Variables	Total number of		The 1st component of MDS for dissimilarity matrix about time profile of cumulative							
	C	D	C	C/P	C/T	CC	D	D/P	D/T	CD
McDonalds (x'_1)	$-0.41^\wedge$	$-0.43^\wedge$	$-0.41^\$$	$-0.38^\$$		$-0.44^\$$	$-0.42^\wedge$	$-0.40^\$$		-0.30
Other fast-food (x'_2)	$0.37^\$$		$0.35^\#$	$0.51^\$$		$0.55^\$$				
Non-fast-Food (x'_3)	0.19		0.24							
Grocery (x'_4)		-0.24					-0.24			
Fitness (x'_5)		$0.84^\wedge$					$0.91^\&$	0.40		
Pharmacy (x'_6)	$-0.45^\$$	$-0.73^\$$	$-0.44^\$$	$-0.47^\#$		$-0.47^\#$	$-0.74^\wedge$	$-0.49^\$$		
Nursing home (x'_7)	$-0.31^\$$	$0.60^\$$	$-0.33^\$$				$-0.61^\$$			-0.31
Low-income (l)	$0.34^\$$	$0.32^\#$	0.32	$0.45^\$$		$0.37^\#$	$0.32^\$$	$0.52^\$$		
Elder (o)		$0.30^\#$			$-0.42^\#$		$0.30^\#$			
Adjusted R-square	0.72	0.73	0.72	0.53	0.13	0.48	0.73	0.50		0.14

All responses were the residuals obtained from fitting the original data versus population (p) and distance to NYC (d)

were associated with more cases and deaths. However, some were not, that is, some were negative while others were positive, e.g., distance to NYC in Table 1 and nursing home in Table 2. It should be noted that different analysis represented different measurements on cases or deaths. The effects of the same factor on different measurement could be different. For example, Column "C" in Table 1 represented the total number of cases, while Column "D/P" meant the proportion of deaths among the populations. The results showed that a larger distance to NYC resulted in more cases, but less death proportion.

Figure 3 showed the clear demarcation of three New Jersey regions, North Jersey, South Jersey and Central Jersey. That was suggested by the Covid-19 cases and deaths data and by the demographics and socioeconomic variables. (this study supports one side of the never ending debate about the existence of Central Jersey!)

5 Conclusions

In this study, we investigated the effects of several socio-economic factors on the number (and number over time) of confirmed Covid-19 cases and Covid-19-related deaths for the twenty one counties of New Jersey during the period March 1st to September 9th, 2020.

We found that counties could be clustered into three groups based on (a) the case totals, (b) the total number of deaths, (c) the time course of the cases and (d) the time course of the deaths. The four groupings were very similar to one another and could all be largely explained by the county population, the percentage of low-income population, and the distance of the county from New York, with the highly populated counties close to New York clustering together as the hardest hit counties and the less populated ones far from it clustering together as the least affected.

We also examined the effects of various food and health factors, where people may gather and potentially result in more cases and found that an increased number of McDonald's, in contrast to other fast-food or non-fast-food restaurants, led to a reduction in the number of cases and deaths regardless of whether or not it was adjusted by the total number of cases or by population. In addition, (a) a larger number of pharmacies resulted in fewer cases and fewer deaths and (b) greater access to fresh food (groceries) led to fewer deaths. However, the number of fitness centers was associated with an increased number of cases and deaths.

Moreover, our study found that these same socio-economic factors could be used to explain the differences in the evolution of the disease between different geographical locations (counties in New Jersey). Furthermore, these factors still emerged as significant even after regressing out the main factors (population and distance to New York).

Overall, we found that the evolution of Covid-19 in New Jersey has been influenced by certain socio-economic factors, which could be helpful for the formulation of public health policies.

Appendix

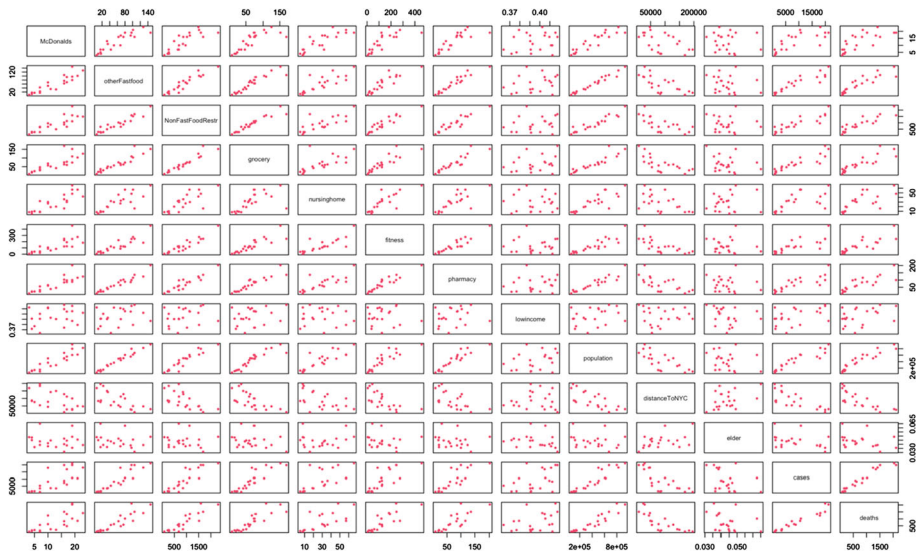## Appendix A: The Matrix of scatter plots

See Figs. 4, 5 and 6.

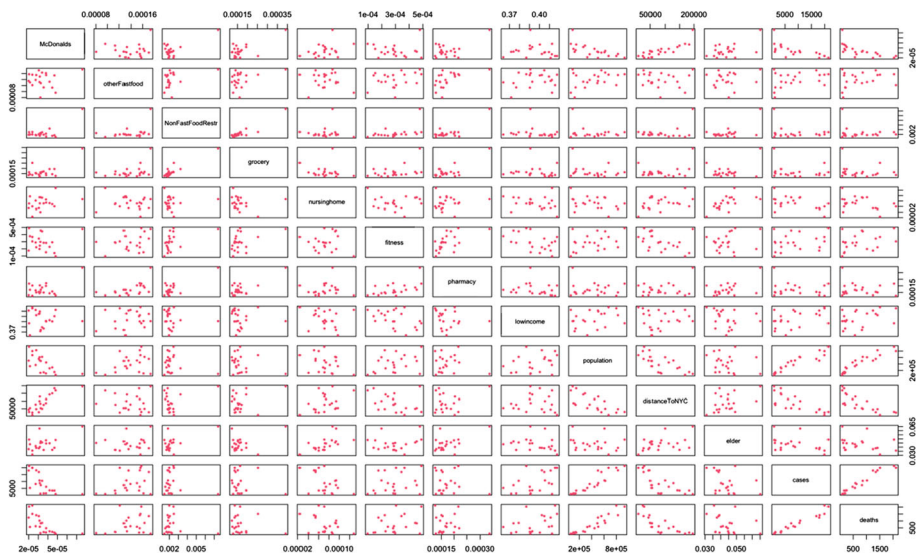**Fig. 4** A matrix of scatter plots for original data

Fig. 5 A matrix of scatter plots for a dataset where food-and-health variables (the first seven variables) were divided by population
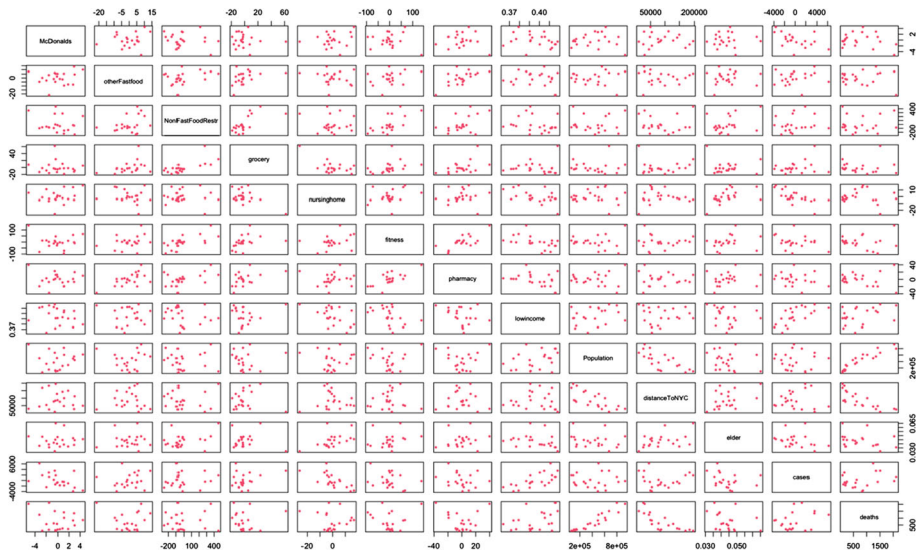
Fig. 6 A matrix of scatter plots for a dataset formulated by residuals. The food-and-health variables, cases and deaths here were the residuals obtained from fitting them vs population and distance to NYC

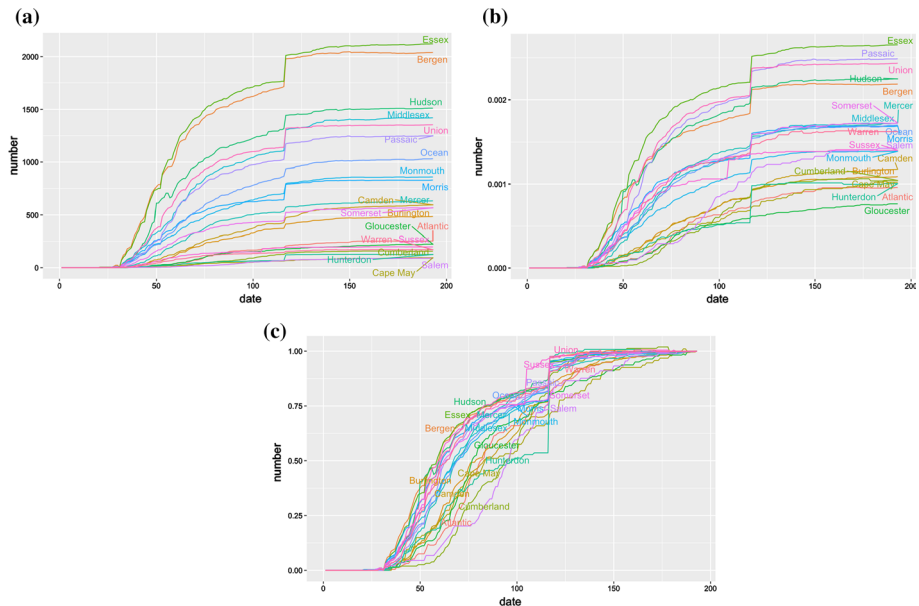## Appendix B: Plots about death time-course data

See Figs. 7 and 8.

Fig. 7 The different time profile deaths data. **a** The raw time-course death data, $\{c_{ij}\}$; **b** the time-course death data divided by population; **c** the time-course death data divided by total number of cases
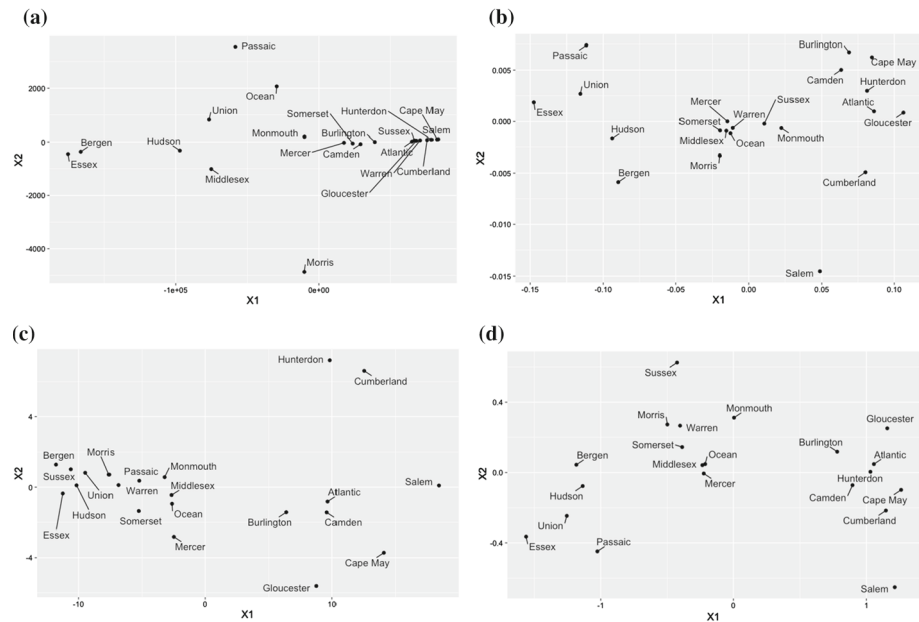
Fig. 8 Scatter plots after Multidimensional scaling for dissimilarity matrix relevant to **a** the raw time-course death data; **b** the time-course death data divided by population; **c** the time-course death data divided by total number of deaths, $\{c_{ij}/c_{mj}\}$; **d** combination

References

2018: ACS 1-year estimates data profiles. https://data.census.gov/cedsci/table?t=IncomeandPoverty&g=0400000US34,34.050000&tid=ACSDP1Y2018.DP03&moe=false&hidePreview=true.

Aztsop, J., & Joy, T. R. (2013). Access to nutritious food, socioeconomic individualism and public health ethics in the USA: A common good approach. *Philosophy, Ethics, and Humanities in Medicine*, *8*(1), 16. https://doi.org/10.1186/1747-5341-8-16.

cc-est2019-agesex-34. https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/counties/asrh/cc-est2019-agesex-34.csv.

Cordes, J., & Castro, M. C. (2020). Spatial analysis of COVID-19 clusters and contextual factors in New York City. *Spatial and Spatio-temporal Epidemiology*, *34*, 100355. https://doi.org/10.1016/J.SSTE.2020.100355.

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infectious Diseases*, *20*(5), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1.

Drefahl, S., Wallace, M., Mussino, E., et al. (2020). Socio-demographic risk factors of COVID-19 deaths in Sweden: A nationwide register study. *Stockholm Research Reports in Demography*. https://doi.org/10.17045/STHLMUNI.12420347.V3.

Gu, T., Mack, J. A., Salvatore, M., et al. (2020). COVID-19 outcomes, risk factors and associations by race: A comprehensive analysis using electronic health records data in Michigan Medicine. *medRxiv*. https://doi.org/10.1101/2020.06.16.20133140.

Hilmers, A., Hilmers, D. C., & Dave, J. (2012). Neighborhood disparities in access to healthy foods and their effects on environmental justice. *American Journal of Public Health*, *102*(9), 1644–1654. https://doi.org/10.2105/AJPH.2012.300865.

Jones, K. E., Patel, N. G., Levy, M. A., et al. (2008). Global trends in emerging infectious diseases. *Nature*, *451*(7181), 990–993. https://doi.org/10.1038/nature06536.

Khalatbari-Soltani, S., Cumming, R. C., Delpierre, C., & Kelly-Irving, M. (2020). Importance of collecting data on socioeconomic determinants from the early stage of the COVID-19 outbreak onwards. *Journal of Epidemiology and Community Health*, *74*(8), 620–623. https://doi.org/10.1136/jech-2020-214297.

Kim, S. J., & Bostwick, W. (2020). Social vulnerability and racial inequality in COVID-19 deaths in Chicago. *Health Education and Behavior*, *47*(4), 509–513. https://doi.org/10.1177/1090198120929677.

Krieger, N. (2001). A glossary for social epidemiology. *Journal of Epidemiology and Community Health*, *55*(10), 693–700. https://doi.org/10.1136/jech.55.10.693.

McLaren, J. (2020). Racial disparity in COVID-19 deaths: Seeking economic roots with census data. https://doi.org/10.3386/w27407

Raisi-Estabragh, Z., McCracken, C., Bethell, M. S., et al. (2020). Greater risk of severe COVID-19 in Black, Asian and Minority Ethnic populations is not explained by cardiometabolic, socioeconomic or behavioural factors, or by 25(OH)-vitamin D status: study of 1326 cases from the UK Biobank. *Journal of Public Health (Bangkok)*, *42*(3), 451–460. https://doi.org/10.1093/pubmed/fdaa095.

Affiliations

Dhammika Amaratunga[1] · Javier Cabrera[2] · Debopriya Ghosh[4] · Michael N. Katehakis[2] · Jin Wang[3] · Wenting Wang[2]

Dhammika Amaratunga
damaratung@yahoo.com

Javier Cabrera
cabrera@stat.rutgers.edu

Debopriya Ghosh
DGhosh22@its.jnj.com

Michael N. Katehakis
mnk@rutgers.edu

Wenting Wang
ww274@business.rutgers.edu

[1] Princeton Data Analytics, Princeton, USA

[2] Rutgers, The State University of New Jersey, New Brunswick, USA

[3] Jinan University, Guangzhou, China

[4] Janssen Research and Development LLC, Raritan, USA