Science of the Total Environment

COVID-19 in New York state: Effects of demographics and air quality on infection and fatality

--Manuscript Draft--

Manuscript Number:	STOTEN-D-21-05500R1
Article Type:	Research Paper
Keywords:	COVID-19; New York State; air quality; PM2.5; Clustering; stepwise regression
Corresponding Author:	Shantanu Sur, Ph. D. Clarkson University Potsdam, NY UNITED STATES
First Author:	Sumona Mondal
Order of Authors:	Sumona Mondal
	Chaya Chaipitakporn
	Vijay Kumar
	Bridget Wangler
	Supraja Gurajala
	Suresh Dhaniyala
	Shantanu Sur, Ph. D.
Abstract:	The coronavirus disease 2019 (COVID-19) has had a global impact that has been unevenly distributed amongst and, even within countries. Multiple demographic and environmental factors have been associated with the risk of COVID-19 spread and fatality, including age, gender, ethnicity, poverty, and air quality among others. However, specific contributions of these factors are yet to be understood. Here, we attempted to explain the variability in infection, death, and fatality rates by understanding the contributions of a few selected factors. We compared the incidence of COVID-19 in New York State (NYS) counties during the first wave of infection and analyzed how different demographic and environmental variables associate with the variation observed across the counties. We observed that infection and death rates, two important COVID-19 metrics, to be highly correlated with both being highest in counties located near New York City, considered as one of the epicenters of the infection in the US. In contrast, disease fatality was found to be highest in a different set of counties despite registering a low infection rate. To investigate this apparent discrepancy, we divided the counties into three clusters based on COVID-19 infection, death rate, or fatality, and compared the differences in the demographic and environmental variables such as ethnicity, age, population density, poverty, temperature, and air quality in each of these clusters. Furthermore, a regression model built on this data reveals PM2.5 and distance from the epicenter are significant risk factors for infection, while disease fatality has a strong association with age and PM2.5. Our results demonstrate that for the NYS, demographic components distinctly associate with specific aspects of COVID-19 burden and also highlight the detrimental impact of poor air quality. These results could help design and direct location-specific control and mitigation strategies.
Response to Reviewers:	The authors have addressed the reviewers comments in the response to reviewers document.

Clarkson

Department of Biology

Shantanu Sur
Associate Professor
ssur@clarkson.edu

Date: July 28, 2021

Dr. Scott C. Sheridan
Associate Editor
Science of the Total Environment

Dear Dr. Sheridan,

We are submitting a revised version of our manuscript entitled “COVID-19 in New York state: Effects of demographics and air quality on infection and fatality” for publication in *Science of the Total Environment*. We are grateful for the thoughtful and extremely valuable comments raised by the reviewers, and we have extensively revised the manuscript taking these comments into consideration. A point-by-point response to the reviewers’ comments is attached. We hope the manuscript is now suitable for publication and await your decision.

Sincerely yours,

Shantanu Sur

COVID-19 in New York state: Effects of demographics and air quality on infection and fatality

Sumona Mondal[a,1], Chaya Chaipitakporn[b,1], Vijay Kumar[a], Bridget Wangler[b], Supraja Gurajala[c], Suresh Dhaniyala[d], and Shantanu Sur[e,*]

[a]*Department of Mathematics, Clarkson University, Potsdam NY, USA*

[b]*David D. Reh School of Business, Clarkson University, Potsdam NY, USA*

[c]*Department of Computer Science, SUNY Potsdam, Potsdam NY, USA*

[d]*Department of Mechanical and Aeronautical Engineering, Clarkson University, Potsdam NY, USA*

[e]*Department of Biology, Clarkson University, Potsdam NY, USA*

[1]*These authors contributed equally to this work.*

**Corresponding author: ssur@clarkson.edu*

Authors' response to reviewers' comments:

We would like to thank the reviewers for the critical and thorough evaluation of our manuscript. The inputs, suggestions, and constructive criticism have helped in addressing many of the weaknesses, improving the quality, and the overall focus of our manuscript. The authors have included here a point-by-point response to the comments and concerns from the reviewers. Specific sections of the comments and concerns are first stated in *italics* and then addressed individually. The changes incorporated in the manuscript to duly address the comments and suggestions are highlighted in yellow and the locations are also mentioned in the individual response.

Reviewer 1

This paper compares C-19 infection and fatality with various demographic and environmental variables. They also identified the relative contribution of the risk variable on C-19 infection and mortality. The authors need to clearly communicate what contribution this paper will make towards new knowledge. Furthermore, thePM2.5 mass concentration values used should be fully motivated and justified. Lastly, the paper needs to communicate how this approach can be applied to other states and how that will enable the implementation of mitigation, and potentially prevention, infections.

[Author Response, AR] The authors appreciate the valuable comments and suggestions made by the reviewer. (1) We have now made extensive revisions throughout the manuscript text that we hope will better communicates the contribution of work towards new knowledge. (2) In the methods and result section we have added the motivation and justification for the use of $PM_{2.5}$ mass concentration values used in this work (page 4, paragraph 1, lines 157-160). (3) We have included in the manuscript text how we envision our analysis and approach could be applied to other states in the US (pages 18-19, paragraph 3, lines 524-535). While the results of our analysis might not be directly applicable towards implementation of mitigation and preventive measures, it could provide important insight in such decision-making processes.

[Reviewer Comment, RC 1] The introduction does not naturally lead the reader to the contribution to new knowledge that this paper claims. It is suggested that this be made clear.

[Author Response, AR 1] We have made substantial edits in the introduction section of the paper to engage the reader towards the new knowledge this paper contributes (pages 2-3, paragraph 3, lines 79-119). Specifically, the fourth paragraph of the introduction is now rewritten to emphasize, (1) why it is necessary to include both demographic and environmental risk variables in the analysis as conducted in the present work, (2) why it is important to consider and compare the impact of risk variables on different aspects of COVID-19 burden, namely infection, death, and fatality, and (3) how the study conducted on New York state helps to address these questions.

2. In the methodology:

[Reviewer Comment, RC 2.1] Line 98: it is said that The data is from March 1, 2020 and May 16, 2020. Does that mean the paper is based on two single days? Or does it cover the period between the two dates?

[Author Response, AR 2.1] We thank the reviewer for pointing out this sentence that indeed created a major confusion suggesting the work was done based on the data form only two single days. The analyses in the present study were performed on COVID-19 cases and deaths that occurred during the period from March 1, 2020 to May 16, 2020, which roughly corresponds with the first wave COVID -19 in the NYS. We have made this correction in the methods section of the manuscript (page 3, paragraph 3, lines 123-125).

[Reviewer Comment, RC 2.2] If the paper is based on 2 dates only, the monthly PM2.5 averages cannot be used. In fact, PM2.5 varies significantly from day to day and I fail to see how one can link two single days with the pollution data. From further reading it does appear that this was a grammatical error and that the study does indeed cover a few months.

[Author Response, AR 2.2] As explained in the previous response, the confusion stemmed from the statement inaccurately describing the period of COVID-19 cases and deaths considered in the study. The analyses performed in the manuscript consider COVID-19 data during the period of March 1, 2020 to May 16, 2020, and has now been corrected in the manuscript text.

[Reviewer Comment, RC 2.3] Line 124: Motivate the averaging of the PM2.5 data from 2000 - 2016 - Is this representative of the PM2.5 mass concentrations during the sampling period?

[Author Response, AR 2.3] We have included the motivation for using averaged $PM_{2.5}$ data from the years 2000 – 2016 in the methods section (page 4, paragraph 1, lines 157-160). In support of this decision, we cited multiple reports that have concluded the historical exposure to $PM_{2.5}$ to be a more relevant variable than the acute exposure in studying the adverse effects on COVID-19 (Wu, X et al., *Sci Adv* 2020, 6 (45), eabd4049; Gupta, A. et al., *Environ Dev Sustain*, 2020, 1-10; Maleki, M. et al., *Environ Re*s, 2021, 195, 110898). Additionally, one study based on NYC have reported that there was no significant change in the air quality after the lockdown when compared with previous years (Zangari, S. et al., *Sci Total Environ*, 2020, 742, 140496). Even if a difference exists between the absolute $PM_{2.5}$ values during the sampling period and the average values computed from the past years, we expect the relative differences of the $PM_{2.5}$ level across different locations of the state to be conserved and therefore, would reflect in its impact on COVID-19.

[Reviewer Comment, RC 2.4] Fig 2 nursing homes are not clear on the map and the quality of the images are poor.

[Author Response, AR 2.4] We have updated the map with a clearer depiction of the nursing home locations on the map. To improve the clarity, we have increased the image size of the map as permissible within the limited figure space. While it might still not be easy to read the names of individual counties in the map without zooming in the document, our primary objective here is

to provide the reader a perspective of the geographical distribution of the counties grouped in each cluster (each cluster is represented by a different color). We also realize that the figure image quality is partially compromised by the lower resolution pdf copy generated during the manuscript submission process.

3. Results:

[Reviewer Comment, RC 3.1] Fig 6 A could do with a better description and explanation of what it provides. That also applies to Fig 6C.

[Author Response, AR 3.1] We thank the reviewer for this suggestion. We have now expanded and rewritten the text pertaining to Figure 6A and Figure 6C to include more details and clarity. The changes are made in pages 14-16, paragraphs 4-5, lines 379-413.

[Reviewer Comment, RC 3.2] It will also be in general useful if the clusters are shortly described in the captions for all figures.

[Author Response, AR 3.2] The authors thank the reviewer for this excellent advice. We have included a short description of the clusters in the captions of all relevant figures and feel this has substantially improved the ease of interpretation of figures.

4. Discussion:

[Reviewer Comment, RC 4.1] In the whole of the document reference is made to deaths, fatality, mortality and these terms seem to be loosely used. I proper definition of what is meant by each term would be really useful

[Author Response, AR 4.1] The authors apologize for the confusion created on using these terms. The terms death rate and mortality rate were used interchangeably to indicate the total number of COVID-19 deaths in a county during the study period per unit population. To be consistent, we are now using only the term "death rate" throughout the manuscript to refer to this metric. The fatality rate of a county was defined by dividing the total number of COVID-19 deaths by total number of COVID-19 infections during the study period. These terms are now defined in the methods section of the manuscript (page 4, paragraph 1, lines 127-132).

[Reviewer Comment, RC 4.2] Line 356: mentions that acute lung inflammation could be a factor contributing to higher susceptibility to the virus. The authors then sate that the average air quality is good. However, the authors missed the point that there are most definitely hot spots across the region and that these concentrations could be significantly higher in some areas. In addition, it is not clear whether the air pollution at the time of the investigated period has been used. Furthermore, the issue of the synergistic effects of increased NO2 levels on pulmonary health should be mentioned - add to that the link between PM mass, ozone and temperature and old age causing cardiovascular episodes.

[Author Response, AR 4.2] We thank the reviewer for this excellent suggestion to incorporate in the discussion section. As advised, we have rewritten the paragraph two of the discussion section

to communicate the relevant points recommended including, (1) the potential contribution of hot spots of poor air quality in the observed effects on COVID-19, (2) the possibility of synergistic effects with other pollutants, especially NO_2 and O_3, and the connection to temperature and old age with predisposition to cardiovascular episodes. These changes are incorporated in pages 16-17, paragraph 3, lines 432-464.

[Reviewer Comment, RC 5] Conclusions: the authors claim that the regression model helped to estimate the relative contribution of the factors in infection - if this is indeed the case – the authors should spell-out these contributions. They conclude with a sentence to say that the understanding the interplay of risk variables can help with developing preventative measures. They need to hypothesis who that would be possible, whether this is of use for any other district and therefore if it is potentially transferable to other states and countries.

[Author Response, AR 5] We thank the reviewer for the critical appraisal of the conclusion section. To address the points raised by the reviewer, we made considerable modifications in the conclusions section (pages 18-19, paragraph 3, lines 509-535). (1) We removed the statement claiming that the regression model helped to estimate the relative contribution of the factors in infection as we realized this to be a rather ambitious statement. Instead, we now state that how the regression models help to identify the major contributors of infection, death, and fatality. (2) We clarified that given the strong influence of anthropogenic factors on environment, both demographic and environmental risk variables are needed to be considered for developing preventative and mitigative measures. (3) We have also described how we envision the analyses could be useful and extended to interpret the impact of these risk factors in other states of the US.

Reviewer 2

The manuscript describes an analysis of what demographic characteristics and environmental variables for each NY county are associated with COVID-19 morbidity and mortality measures. However, there are several concerns with the manuscript, and places in the text where more information is needed to understand what specifically was done in the analysis, that should be addressed before considering it for publication.

[Author Response, AR 1] We sincerely thank the reviewer for providing a critical review of the manuscript. We have made an extensive revision to the manuscript text to address the concerns and suggestions made by the reviewer. The details of the revisions made to the manuscript are provided in the response to individual comments below. We feel these changes have helped the manuscript to better communicate the information, and also substantially improved the overall quality of the manuscript.

[Reviewer Comment, RC 1] The introduction provides a very general overview of the COVID-19 pandemic and what happened around the world with regard to it. However, why haven't you described the situation in NY? I strongly suggest you rewrite the Introduction to include a more

detailed description of the disease occurrence in New York State and NYC and use that as a justification for the analysis you conducted. Without this, the introduction is lacking.

[Author Response, AR 1] We thank the reviewer for the constructive criticism on the content of the introduction. As suggested, we have made considerable modifications in the introduction section to provide a context of the current study, which is based on New York State data. Specifically, we have rewritten the fourth paragraph and of the introduction added a fifth paragraph to describe the situation in NY State and NY City and how the study on this state is suitable for the research questions we have asked in this manuscript. These changes are included in the pages 2-3, paragraph 3, lines 79-119.

[Reviewer Comment, RC 2] Introduction Line 89 - "Using publicly available data, we grouped the counties into clusters based on ….". Here and throughout the methods section the analysis plan is just not clear. Was the unit of observation in the regression analysis a county for the whole time period, a county for each week of the study period, a cluster of counties for each week of the study period? Clarify that here, but also make it clearer in the methods section as well.

[Author Response, AR 2] We apologize for the lack of clarity in communicating the analysis plan. In this work, we have taken two approaches for analysis: (1) Compute the infection, death, or fatality rates of the counties during the study period, group the counties into clusters based on these variables, and then investigate the association of the clusters with various demographic and environmental factors. (2) Regression analysis where infection, death, or fatality rates of individual counties during the study period were considered as response variables and the measures of demographic and environmental variables were used as predictor variables. Together these two approaches help to identify the association of the risk factors with different aspects of COVID-19 burden (infection/death/fatality) and then identify the major contributor to a specific burden. We have clarified the analysis plan in the introduction section (page 3, paragraph 2, lines 109-119), and also considerably revised the methods section (details of these modifications are provided in the author response to address some of the later comments from the reviewer) to add clarity.

[Reviewer Comment, RC 3] Line 115 - Why was Manhattan chosen as the epicenter? What about the other NYC boroughs? Provide a reference to support this.

[Author Response, AR 3] The reviewer is correct in that the NYC as a whole was considered as the epicenter of the COVID-19 outbreak without further resolving into individual boroughs (Wadhera, R. K. et at., *JAMA*, 2020, 323(21), 2192-2195; Reichberg, S. B. et al., *Clinical Infectious Diseases*, 2020, 71(12), 3204-3213.). In this work we have chosen Manhattan over other boroughs as the disease epicenter due to its central location in the NYC. We have clarified this in the manuscript and also added references to support NYC as the epicenter of the outbreak. These changes made in page 4, paragraph 1, lines 142-146.

[Reviewer Comment, RC 4] Line 124 - "…temporally averaged PM2.5 data was used in the study". Provide the unit of time on which it was averaged. Were they weekly averages? Daily?

[Author Response, AR 4] We have used a single $PA_{2.5}$ data in the analysis that was obtained by temporally averaging monthly data for the years of 2000 – 2016, as described before by Wu, X., et al. (*Science Advances*, 2020, 6(45), eabd4049). This average value used as a measure of chronic exposure to $PA_{2.5}$. This is now clarified in the methods section (page 4, paragraph 1, lines 148-157).

[Reviewer Comment, RC 5] Methods text starting with 2.2 Statistical Analyses. As you describe each section of the analysis, please make it clear at the end of that paragraph, what is the product of that portion of the analysis, and then how it is used in the next section or in the overall analysis. As a reader, I was not able to follow the path of the statistical analysis and how each section (e.g., K-means clustering) was used in the next section.

[Author Response, AR 5] We thank the reviewer for the critical feedback to improve the methods section. We have made substantial edits to the descriptions of Statistical Analysis to improve the clarity of communication. As advised, at end of each technique we mentioned how the method is used in the subsequent section or in the overall analysis. These changes are made in the text describing k-means clustering (page 4, paragraph 2, lines 164-167; page 5, paragraph 1, lines 177-180) and throughout all subsections in 2.2 Statistical Analysis.

[Reviewer Comment, RC 6] Figure 1 - Pictures are too small to detail the 3 clusters. Further, are the clusters just NYC, counties surrounding NYC, and then upstate NYS (i.e., the rest of NY), or something very similar to that?

[Author Response, AR 6] We agree to the reviewer that larger pictures would have made it easier for the reader to view the finer details. We have made an effort to increase the size of the maps, however, their sizes are still limited by the allowed width of the manuscript. We would like to mention that to overcome this limitation, we color coded the clusters in each category. Further, the distribution of values (infection or death rates) in a cluster is shown in bar plots and the maps depict the spatial distribution of the counties in each cluster. With this approach we wanted to convey to the reader the following information about NYS counties in an efficient manner: (1) Absolute estimates of infection/death rates along with the range of values for each cluster, (2) Number of counties in each cluster, and (3) the geographical distribution of counties belonging to each cluster. We would like to emphasize that our primary objective for the figure is to provide the reader a comprehensive understanding of the cluster characteristics and not drawing attention regarding the value or location of individual counties in the plot or the map, respectively.

Regarding the query on what the clusters represent, the clusters were generated using k-means clustering technique and are based on infection or death rates of the counties. The range of values of these parameters in each cluster could be visually checked in Figure 1A and 1B. Incidentally (as the reviewer has noticed), we found that the counties in cluster 1 (high infection/death) and cluster 2 (intermediate infection/death) were primarily located in the NYC

and counties surrounding NYC, respectively, although some finer difference could be observed between the categories for infection and death (Figure 1C). Cluster 3 (low infection/death) consisted of mostly the upstate counties.

[Reviewer Comment, RC 7] Line 158 - provide the Bonferroni corrected p-value(s).

[Author Response, AR 7] We have included the Bonferroni corrected p-value (page 7, paragraph 1, lines 191-192).

[Reviewer Comment, RC 8] Lines 172-180 - provide the output/effect estimate from the ARIMA model. Do you only have 3 rates (and 95% CIs) for each of the 3 clusters? Again, missing details on what this model produces and then how you used it to make inference.

[Author Response, AR 8] As advised, we have included the output/effect-estimate from the ARIMA model and explained how this model is applied to make inference in the analysis (page 7, paragraph 1, lines 194-199 and paragraph 3, lines 209-215).

Since the number of Environmental Protection Agency (EPA) monitoring sites are limited across the NYS, we have selected one representative EPA site for each cluster in each category (infection/death/fatality). Thus, we have three predicted rates (and 95% Cis) for each of the three clusters for a given category. This information is provided in the results section where ARIMA analysis is discussed (page 13, paragraph 1, lines 334-340).

[Reviewer Comment, RC 9] Lines 189-199 - Same issue with the model. A clearer description of the model and what outcome and effect estimate(s) is/are produced by the model.

[Author Response, AR 9] We have substantially modified the "Regression Models" subsection in the methods to make a clearer description of the model, the outcome and effect estimates, and how the model is used to analyze the data in the manuscript. These changes are made in the following locations: page 7, paragraph 4, lines 218-221 and 223-224; page 9, paragraph 1, lines 226-239.

[Reviewer Comment, RC 10] Line 230 - I am concerned about ecologic fallacy here and taking an association with or within a county and assigning it to all those with that characteristic in the county. Please discuss this or provide an argument on why it is or is not an issue here.

[Author Response, AR 10] Thank you for raising the concern about the ecological fallacy. We would like to mention that here and throughout the manuscript, our analyses are based on summary data from individual counties (e.g., COVID-19 infections, deaths, age, ethnicity, etc.). This poses a limitation to draw any inference on an individual in the population; therefore, based on the information available, our attempt is to identify and report the association of variables at the county level and refrain from assigning it to individuals in the county having those variable characteristics. For example, we studied how infections and death rates associate in the counties in the NY State, or how higher percentage of population with advanced age associate with

infection, death, and fatality statistics of the counties. Even though the analyses do not provide the inference at the resolution of an individual person, we believe it could still provide meaningful insight at the county level.

In line 230, we have shown the relationship between the fatality rate and infection rate among NY State counties. Since we observed no correlation between these two variables (we found counties with high fatality/low infection, low fatality/high infection, and low fatality/low infection but not with high fatality/high infection), this motivated us to hypothesize that for the counties, the risk factors associated with infection and fatality would be different. In subsequent section, we explored this possibility at a greater detail by clustering the counties.

[Reviewer Comment, RC 11] Figure 4 - I do not understand what is being presented here. Please clarify this in the text and describe more clearly what inference was made based on what statistic(s) or effect estimates shown in these figures.

[Author Response, AR 11] We thank the reviewer for pointing out the lack of clarity in communicating the results presented in the Figure 4. We have edited the figure caption and made considerable changes in the corresponding section of the manuscript text (page 13, paragraph 1, lines 330-354) to clearly communicate the inference and how the results (effect estimates) in this figure help to attain that inference.

[Reviewer Comment, RC12] Discussion line 343 - "We observed that infection, death, and fatality rates have a significant association with air quality and various demographic factor. "Please remove the emphasis on the significance of the association and instead focus on the size and direction of the effect on which you base your conclusions. Here and throughout the discussion, provide the quantified effect estimates on which you base your conclusions, so the reader can more easily determine if they agree with you. The significance of the association is far less meaningful.

[Author Response, AR 12] The authors appreciate this valuable suggestion by the reviewer. While the authors believe that the tests for significance is valuable to determine whether an observed association is real or not, we agree with the reviewer that it contributes far less in providing meaningful insight. As advised, we have removed the emphasis on the significance of association and provided more details on the direction of and size of the effects. Additionally, the manuscript is revised to include quantified effect estimates wherever appropriate. These changes are made in the first paragraph of discussion section (page 16, paragraph 2, lines 418-425), and throughout the discussion and results sections.

ABSTRACT

The coronavirus disease 2019 (COVID-19) has had a global impact that has been unevenly distributed amongst and, even within countries. Multiple demographic and environmental factors have been associated with the risk of COVID-19 spread and fatality, including age, gender, ethnicity, poverty, and air quality among others. However, specific contributions of these factors are yet to be understood. Here, we attempted to explain the variability in infection, death, and fatality rates by understanding the contributions of a few selected factors. We compared the incidence of COVID-19 in New York State (NYS) counties during the first wave of infection and analyzed how different demographic and environmental variables associate with the variation observed across the counties. We observed that infection and death rates, two important COVID-19 metrics, to be highly correlated with both being highest in counties located near New York City, considered as one of the epicenters of the infection in the US. In contrast, disease fatality was found to be highest in a different set of counties despite registering a low infection rate. To investigate this apparent discrepancy, we divided the counties into three clusters based on COVID-19 infection, death rate, or fatality, and compared the differences in the demographic and environmental variables such as ethnicity, age, population density, poverty, temperature, and air quality in each of these clusters. Furthermore, a regression model built on this data reveals $PM_{2.5}$ and distance from the epicenter are significant risk factors for infection, while disease fatality has a strong association with age and $PM_{2.5}$. Our results demonstrate that for the NYS, demographic components distinctly associate with specific aspects of COVID-19 burden and also highlight the detrimental impact of poor air quality. These results could help design and direct location-specific control and mitigation strategies.

Keywords: COVID-19, New York State, air quality, $PM_{2.5}$, clustering, stepwise regression

1 Introduction

The impact of the COVID-19 pandemic on global health and economy has exceeded well over the severity of any other communicable diseases in recent history (Baldwin and Di Mauro, 2020; Sarkodie and Owusu, 2020a). The pandemic has also stimulated and significantly accelerated global research into coronaviruses, airborne disease transmissions, and development of new vaccines. Within a short span of time, scientists have succeeded in obtaining critical information on the structure and genomic sequence of the virus pathogen SARS-CoV-2, mechanism of virus infection to host, modes of transmission, and injury to host organs induced by the virus. The research findings have accelerated the development of vaccines and established preventive measures such as the use of masks. Simultaneously, there has been a significant effort to understand the association of COVID-19 to demographic and environmental factors, to explain the geographical or seasonal variability in disease burden (Goldstein and Lee, 2020; Karmakar et al., 2021; Perone, 2021; Sorci et al., 2020). Underscoring precise influences of human demographics and environmental factors on the pandemic would be important toward developing effective public health and social measures.

Among the demographic variables, age, gender, ethnicity, and population density are reported to impact COVID-19. Advanced age is shown to significantly increase the fatality from COVID-19. A study conducted on hospitalized patients in the New York City (NYC) area found 84% of the total deaths occurred in people aged above 60 years (Mesas et al., 2020; Richardson et al., 2020). Moreover, males were seen to be more susceptible to suffer from COVID-19 complications and fatality (Pradhan and Olsson, 2020). Although the mechanism underlying such predisposition of age and sex is not completely understood, the presence of preexisting health conditions and a lowered immunity associated with higher age are thought to be two major factors (Mesas et al., 2020; Pradhan and Olsson, 2020; Richardson et al., 2020). Chronic comorbidities such as hypertension, ischemic heart disease, diabetes, and chronic obstructive pulmonary disease (COPD) are more common in older age and poses risk for severe outcomes (Lusignan et al., 2020; Richardson et al., 2020). Studies focused on the impact of COVID-19 on the ethnic composition also revealed vulnerabilities of certain ethnicities to the disease. In the US, a disproportionately higher number of COVID-19 infections and deaths are observed among African Americans and Hispanic Americans relative to their share of population (Martinez et al., 2020; Yancy, 2020). Socioeconomic disparities leading to increased exposure and lower access to healthcare are thought to contribute to such vulnerability. High population density is reported to increase the risk of COVID-19 spread (Arif and Sengupta, 2020; Copiello and Grillenzoni, 2020), although it is not the sole determining factor as many dense metropolitan cities in Japan, South Korea, China, and Singapore have observed a low infection rate (Lee et al., 2020; Rocklöv and Sjödin, 2020).

The association of environmental factors such as air quality and meteorological parameters to the adverse effects of COVID-19 has been investigated in multiple studies. Air pollution is of particular interest as chronic exposure to air pollutants is linked to multiple chronic respiratory and cardiovascular diseases such as COPD, ischemic heart disease, and hypertension—diseases which are known to increase COVID-19 fatality (Feng et al., 2016b; Guan et al., 2016; Wellenius et al., 2012). Additionally, air pollution substantially increases the risk of respiratory infections including viral infections (Chauhan and Johnston, 2003; Feng et al., 2016a). Fine particulate matter in the air, especially $PM_{2.5}$ (particulate matter with aerodynamic diameter 2.5 µm or less) has been linked to many of these pollution-mediated health effects (Brook et al., 2010; Hopke et al., 2019; Xing et al., 2016). Early reports indicate a positive association of $PM_{2.5}$ with both COVID-19 transmission and fatality (Gupta et al., 2020; Lolli et al., 2020; Pozzer et al., 2020; Wu et al., 2020). Analysis of meteorological factors based on the data from 30 Chinese cities revealed low temperature, less diurnal temperature variation, and low humidity favor the transmission of COVID-19 infection (Liu et al., 2020). This finding was supported by a larger-scale study using data from the top 20 countries with infections, and further claimed low wind speed, surface pressure, and precipitation to increase the risk of disease spread (Sarkodie and Owusu, 2020b).

While the connection of COVID-19 with demographic and environmental factors has been demonstrated by multiple studies, majority of these studies focused on disease transmission or disease fatality alone, and the analyses were directed to either the demographic or the environmental variables. Since anthropogenic factors have a substantial impact on the environmental variables, consideration of both demographic and environmental factors in the

analysis is expected to increase the robustness of inference and reduce the risk of any spurious association (Copiello and Grillenzoni, 2020). Collating information from existing studies to understand the relative impact of these risk factors on infection burden and disease fatality is challenging since these studies were conducted in different geographical locations, and multiple additional confounding factors such as testing and screening strategies, healthcare infrastructure, and socio-cultural practices could contribute to the wide variability of COVID-19 infection and fatality observed across countries or even between different regions within a country (Auger et al., 2020; Chen and Krieger, 2021; Miller et al., 2020). Therefore, to assess the influence of both demographic and environmental factors on COVID-19, ideally the data should be from a geographical location where these factors show considerable variation with low variability of other confounding factors. New York State (NYS), located in the USA fits well as a potential location to conduct such study as it offers a wide range of variation in its demographic landscapes with urban, population-dense, ethnically diverse counties near NYC to many rural, white-dominated, population-sparse counties located in the upstate region. The $PM_{2.5}$ distribution across the state demonstrates a consistent pattern with distinct variation across regions (Jin et al., 2019). Additionally, state-wide implemented policies, including public health measures, and hospital care would help to reduce the potential differences due to the confounding factors mentioned above. Analyses conducted in the NYC area revealed that multiple demographic factors such as ethnicity, male gender, poverty, and household crowding are associated with increased COVID-19 infection, hospitalization, or death (Chen and Krieger, 2021; Reichberg et al., 2020). Studies from NYC metropolitan area also suggested a potential connection of air quality and meteorological variables such as temperature and humidity to higher COVID-19 transmission (Adhikari and Yin, 2020; Bashir et al., 2020). Inclusion of data from the entire NYS is expected to capture a wider variability of these variables and provide a deeper insight into their role in the COVID-19 burden.

In this work, we considered the NYS data at county-level resolution and attempted to relate the variability in infection, death, and fatality with selected demographic and environmental factors during the first COVID-19 wave. Using publicly available data, we first grouped the counties into clusters based on COVID-19 infection, death, or fatality rates during the study period, and investigated the association of these clusters with various demographic factors (e.g., population density, the proportion of African American and Hispanic American population) and environmental factors (e.g., $PM_{2.5}$ and temperature). To identify the risk variables that have major contributions on specific aspects of COVID-19 burden, regression models were then built using data from individual counties, where infection, death, or fatality rates were considered as response variables while demographic and environmental factors were used as predictor variables.

2 Methods

2.1 Study Area, Data Source, and Variables

For this study, COVID-19 infection and death count during the period of March 1, 2020, to May 16, 2020, were obtained for all 62 counties in the NYS from publicly accessible information available at Syracuse.com. The population estimates for each county were obtained from the

2018 US Census Bureau's American Community Survey (ACS) website (https://www.census.gov/programs-surveys/acs)(U.S. Census Bureau, 2018). Infection and death rates from COVID-19 for each county were calculated by dividing the cumulative infection and cumulative death counts during the study period by the total population of the county, and expressed as number per 100,000 population. The fatality rate of a county was obtained by dividing the cumulative death count by cumulative infection count during the study period and presented as the number of deaths per 10,000 infected population. In addition to the total population, the ACS census database was used to collect the following information for each county:(1) Area; (2) population with age ≥ 55 years; (3) poverty levels; (4) Hispanic American population (Martinez et al., 2020); and (5) African American population (Yancy, 2020). From this information (1) population density (population/square mile), (2) proportion of the population with ≥ 55 years (expressed as %), (3) proportion of Hispanic American (expressed as %), and (4) proportion of African American (expressed as %) population was calculated for each county. All factors except population density and distance from the epicenter were converted to percentages by county. The nursing home locations across the NYS counties were obtained from the Department of Health and Human Services. The data was retrieved through ArcGIS Map 10.7.1 (Monmonier and Giordano, 1998). The distance of a county from Manhattan, located at the center of NYC (considered as the disease epicenter (Reichberg et al., 2020; Wadhera et al., 2020)), was used as the distance of the county from the disease epicenter and was calculated by measuring the distance between the centroids of two locations using ArcGIS Map 10.7.1 software. The temperature and Air Quality Index (AQI) information were obtained from Environmental Protection Agency (EPA) measurements available through the United States EPA website (http://www.epa.gov/ttn/airs/aqsdatamart). Hourly outdoor temperature and daily AQI data collected by EPA over a span of 5 years (2015-2019) were used in this study. For county-level $PM_{2.5}$ estimates, temporally averaged $PM_{2.5}$ data previously published by Wu et al. (Wu et al., 2020) were used in this study. Briefly, the monthly averages of $PM_{2.5}$ estimates over the entire US were made at 0.1^{o} X 0.1^{o} grid resolution through a combination of satellite-derived estimates, ground-based measurements, and their statistical fusion through a geographically weighted regression model (Donkelaar et al., 2019). This data was further aggregated to the geographical confinement of a county and temporally averaged for the years 2000 – 2016 to obtain a single $PM_{2.5}$ estimate for each county (Wu et al., 2020). We used this average $PM_{2.5}$ data from the past years in the current study. While the exact mechanisms by which $PM_{2.5}$ influences COVID-19 are not fully understood yet, our choice of average $PM_{2.5}$ from past years is motivated by the findings of multiple studies that point to the association of historical exposure of $PM_{2.5}$ to the disease (Gupta et al., 2020; Maleki et al., 2021; Wu et al., 2020).

2.2 Statistical Analyses

2.2.1 K-Means Clustering:

The counties in the NYS were classified into three categories using k-means clustering technique. Partitioning the counties into three disjoint clusters on the basis of infection, death, and fatality was performed to explore any common pattern that might exist among the counties classified within a cluster. For the implementation of the clustering algorithm, the value of k was

set in advance along with the assignment of initial centroid positions for the clusters (Fahim et al., 2006). The algorithm started with the random initialization of the positions of centroids and was followed by two steps. The first step assigned each sample to its nearest centroid. The second step created a new centroid by taking the mean value of all the samples assigned to each previous centroid. The differences between the old and the new centroids were computed, and the algorithm repeated these last two steps until this difference was less than a threshold. The model used Euclidean distance for the calculation of the distance and the threshold considered was 0.0001. In the end, the centroids were fixed, did not move anymore, signifying the convergence criterion for clustering, and resulted in three distinct clusters. Clustering for infection and death was performed using the infection and death rate values from each county. To cluster the counties for fatality, k-means clustering technique was implemented on infection and fatality rate, considering them as two dimensions. Clusters of counties constructed this way were used to study the association with demographic and environmental risk factors.

Table 1. Publicly available data sources used in this study.

Data	Source
Covid-19 cases & deaths	Coronavirus in NY: Cases, maps, charts, and resources (https://www.syracuse.com/coronavirus-ny/)
Population estimates & demographics 2018	US Census Bureau's American Community Survey (ACS) (https://www.census.gov/programs-surveys/acs)
Temperature & air quality index	EPA (http://www.epa.gov/ttn/airs/aqsdatamart)
Nursing homes locations	The Department of Health and Human Services (HHS) (https://www.arcgis.com/home/item.html?id=b3813b2d3a054c378247bf32bcd8d203)
Satellite $PM_{2.5}$ estimates	Air pollution and COVID-19 mortality in the United States, Harvard University (http://github.com/wxwx1993/PM_COVID)

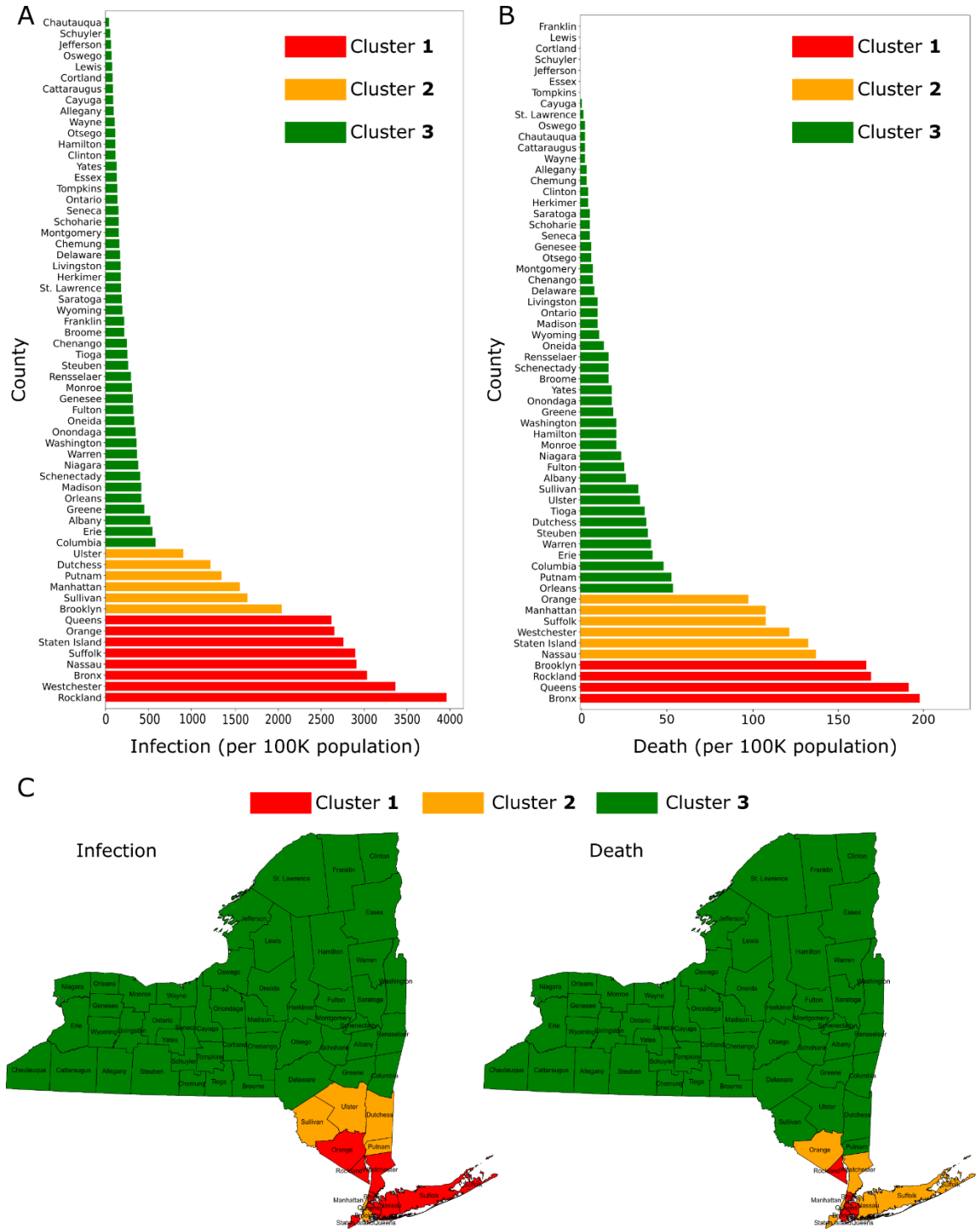


Figure 1. Infection and death from COVID-19 in NYS counties (data till May 16, 2020). (A-B) Infection rates (A) and death rates (B) in individual counties, which are further grouped into three clusters using k-means clustering technique. (Clusters **1, 2,** and **3** represent the counties with high, intermediate, and low infection or death) (C) Maps of NYS showing the locations of counties in each cluster.

184

2.2.2 *Tests for Significance:*

Statistical comparisons of demographic and environmental variables between the clusters were performed using Kruskal-Wallis (KW) test, a non-parametric equivalent of one-way analysis of variance (ANOVA), since the data were non-normally distributed. Once the KW test statistic was found to be significant, multiple comparisons were conducted using Mann-Whitney U test after making the Bonferroni corrections. All analyses used 2-sided statistical tests and $P < 0.10$ was considered as significant. The Bonferroni correction was set at the significance cut-off value of 0.03.

2.2.3 *Autoregressive Integrated Moving Average (ARIMA) Model:*

Temperature and AQI time series data from EPA were used to build ARIMA models to obtain predicted estimates of these variables. Data from one representative EPA site for each of the three clusters in each category of infection, death, and fatality were included in the analysis. The models were constructed using time series data from the years 2015-2019. Hourly outdoor temperature data and daily AQI data collected from EPA were first converted to weekly data before using in the model.

In ARIMA model, the future values of a variable are predicted by a linear combination of past values and errors (Hyndman and Athanasopoulos, 2018). The model is often expressed as ARIMA (p, d, q), where p, d, and q represent the order of auto-regression, the degree of trend difference, and the order of moving average, respectively. The model is essentially a combination of three parts: (1) The first part is the auto-regressive model, which uses the linear combination of past values of the variable to forecast the next value and is referred as an AR(p) model, an autoregressive model of order p. (2) The second part is the integrated (I), which is computed by taking the difference between the consecutive observations to make the data stationary. (3) The third part is the moving average (MA) model, referred as MA(q) and equivalent to a regression model that involves past forecast errors as predictors. Augmented Dickey-Fuller (ADF) unit-root test was performed prior to model building to confirm the stationarity of each time series data. Implementing the ARIMA model, predicted time series values with 95% confidence interval were determined for each condition. The model goodness of fit was further evaluated by calculating the Akaike information criterion (AIC). Model predicted values for each cluster within a category were used to compare the temporal pattern of temperature and AQI between the clusters.

2.2.4 *Regression Models:*

Regression models were built using the data from individual counties of NYS. Three separate models were built where infection, death, and fatality rate were considered as the response variable, while demographic and environmental factors were used as predictor variables for all three models. Variables were first evaluated for normality of distribution by visual inspection of histograms followed by the Shapiro-Wilks test for normality. The univariate method of outlier detection was used to eliminate outliers in the predictors. Correlations between variables were examined by calculating

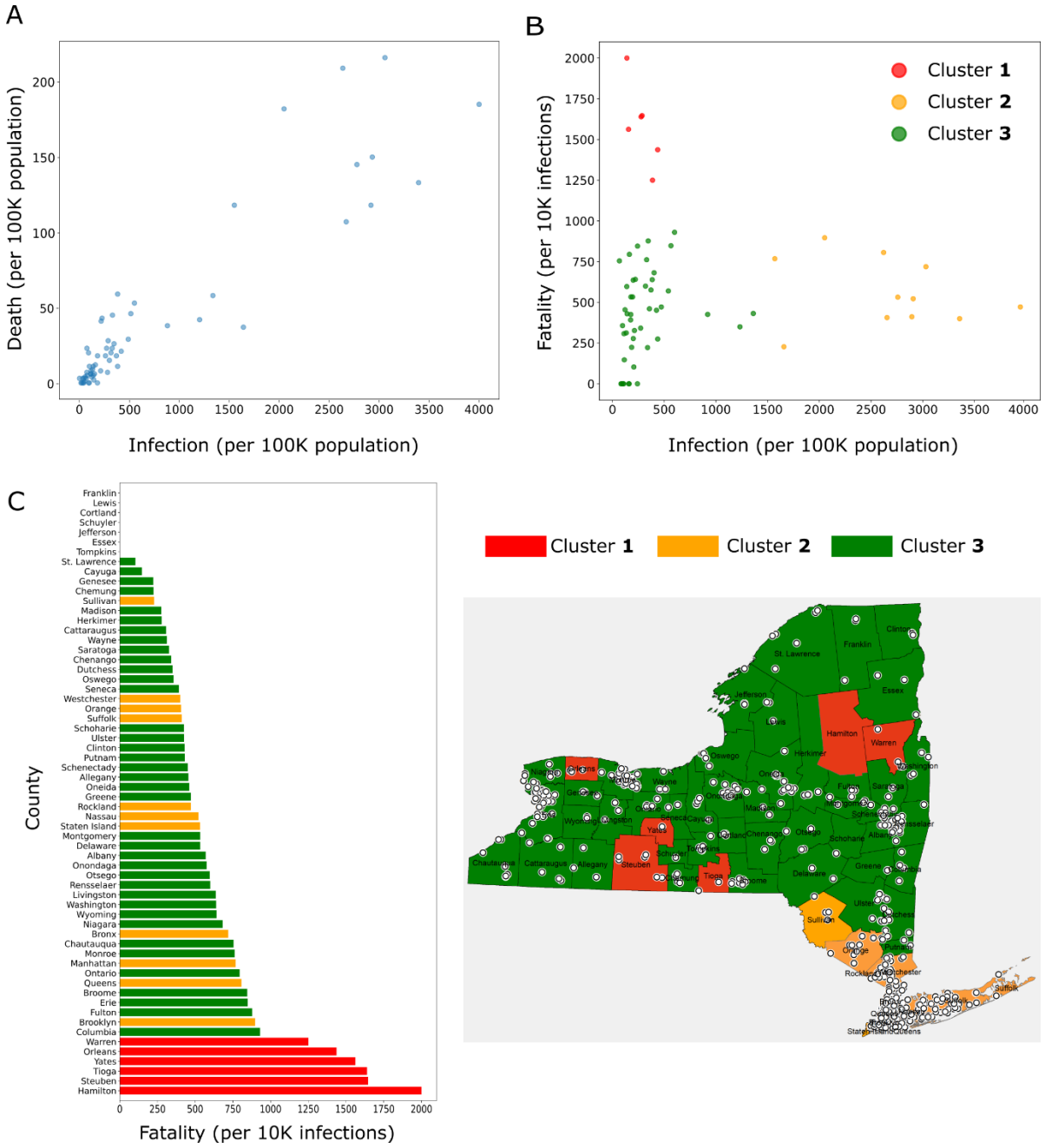


Figure 2. (A) Scatter plot showing the relationship between COVID-19 infection and death rates in NYS counties. (B) Fatality rate is plotted against infection rate for each county; the counties are further grouped for fatality into 3 clusters using two-dimensional k-means clustering method. Cluster **1** includes counties with high fatality and low infection; cluster **2** includes counties with low fatality and high infection; cluster **3** includes counties with low fatality and low infection. (C) Fatality rates of the counties with the clusters indicated by color *(left);* map of NYS showing the locations of counties belonging to each cluster *(right).* Locations of nursing homes are also depicted in the map by white circles.

225

Pearson's correlation coefficients between the predictor and response variables. Multicollinearity between the predictor variables was further examined by computing variance inflation factor (VIF), which measures the inflation in the variances of parameter estimates due to multicollinearity. An upper cut-off value of VIF was set as 5 to minimize the contribution of multicollinearity in our model (Chatterjee and Simonoff, 2013). A stepwise forward selection procedure was implemented to evaluate the contribution of predictor variables in infection, death, and fatality from COVID-19. The forward selection algorithm for stepwise regression starts with an empty model with predictor variables are added sequentially along with the measurement of model accuracy. This process is repeated until all variables are incorporated into the model. The residuals of the regression models were checked for model adequacy and outliers were removed when needed. The goodness of the model is interpreted by the adjusted R^2 value and the contribution of an individual variable is assessed from the order in which the variable was entered in the model. P values of the regression coefficients of the predictor variables were used to assess if their incorporation made a meaningful addition to the model.

All analyses were performed using version 3.6.9 of the Python programming language.

3 Results

3.1 Distribution of COVID-19 in NYS counties

The infections and deaths from COVID-19 in the NYS between March 1 and May 16, 2020, were considered in our analysis. This time window roughly corresponds to the first COVID-19 wave observed in the NYS. To understand the distribution of infections and deaths across the counties within the state, we grouped the counties into three clusters based on each of these variables. Infection and death rates were calculated for all 62 counties in the NYS and then the counties were classified into three clusters using k-means clustering technique. Cluster **1** included counties with a high rate of infection or death, cluster **3** incorporated counties with a low rate, and in cluster **2** the rates were intermediate between the other two clusters (Figure 1A, 1B). For infections, we observed that the cluster **1**, where the infection numbers ranged 2,500-4,000 per 100,000 population, consisted of 8 counties (Rockland, Westchester, Bronx, Nassau, Suffolk, Staten Island, Orange, and Queens), all located in close proximity within the downstate NY (Figure 1C; counties shaded in red). Cluster **2** was formed by 6 counties located near to cluster **1**, namely Ulster, Dutchess, Putnam, Manhattan, Sullivan, and Brooklyn (Figure 1C; counties shaded in yellow). The counties of upstate NY fell in the cluster **3** (Figure 1C; counties shaded in green) where the infection rate was <500 per 100,000 population, well below the other two clusters. The clusters for COVID-19 death showed a similar distribution to infection. Four counties in downstate NY, namely Bronx, Queens, Rockland, and Brooklyn were included in the cluster **1** with death rates ranging from 175 to 200 per 100,000 population (Figure 1B, 1C; counties shaded in red). Of the remaining counties, 6 neighboring counties belonged to cluster **2** (counties shaded in yellow), and the rest of upstate counties were included in the cluster **3** (counties shaded in green) with a death rate <50 per 100,000 population. Thus, clustering the counties followed by visual inspection revealed that higher COVID-19 infections and deaths were from the counties located in downstate NY (Figure 1C).

A similar pattern in the distribution of counties in the clusters for COVID-19 infection and death suggests an association between these two variables, which was confirmed from the scatter plot

(Figure 2A) and a strong positive correlation (Pearson's correlation; r = 0.92, P < 0.0001). The observation suggests that the number of infections in a county is a key determinant for the number of COVID-19 deaths.

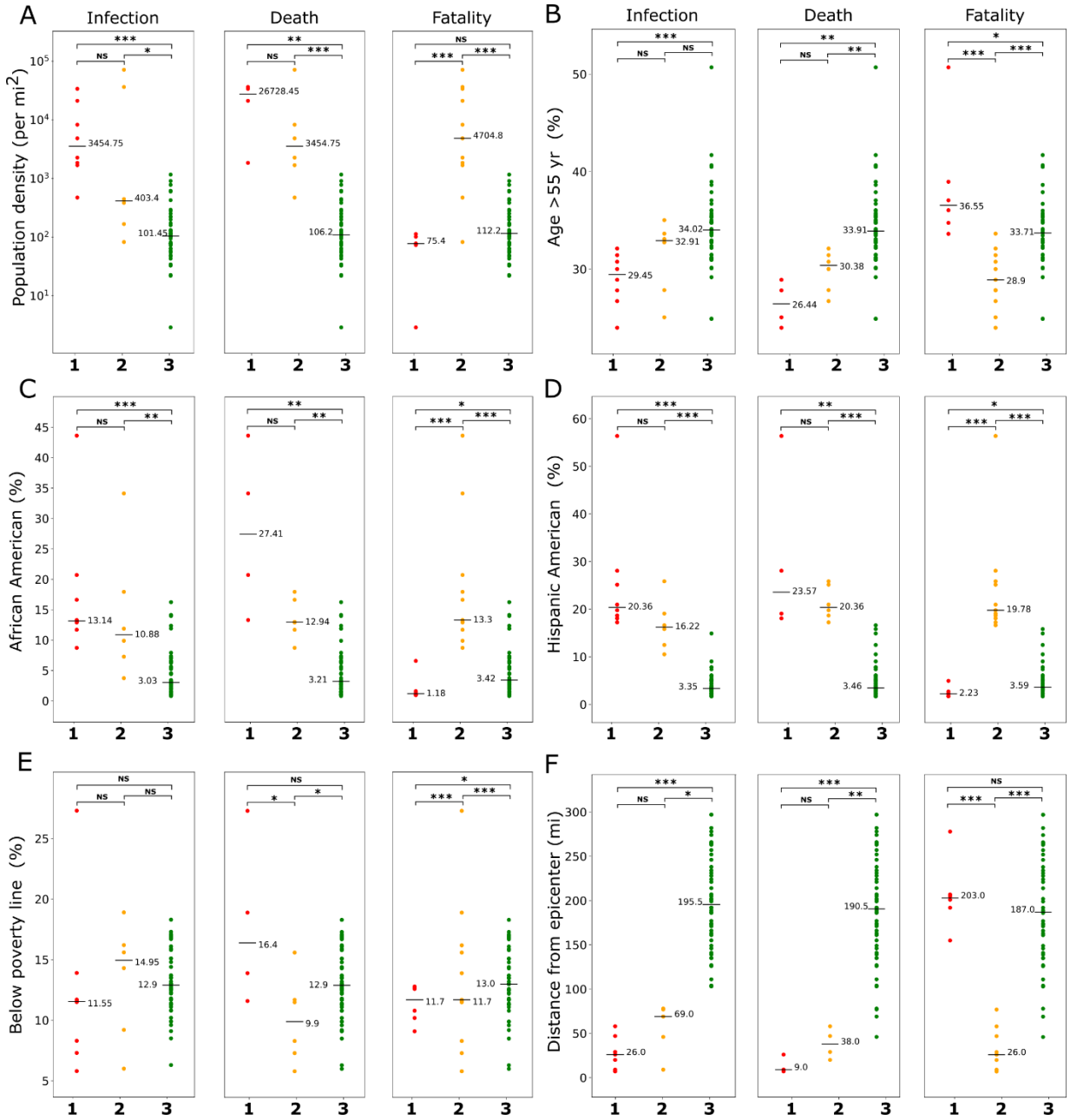


Figure 3. Association of demographic variables with COVID-19. Counties grouped into clusters by infection, death, or fatality were compared for (A) population density, (B) age (> 55 yr), (C) African American population, (D) Hispanic American population, (E) population below poverty line, and (F) distance from epicenter. Horizontal bars represent medians. ***P < 0.01, **P < 0.05, *P < 0.10, NS not significant (Kruskal-Wallis test followed by Mann-Whitney U test with Bonferroni corrections). For infection and death category, clusters **1**, **2**, and **3** represent high, intermediate, and low rates of infection or death; for fatality category,

cluster **1** indicates high fatality and low infection, cluster **2** indicates low fatality and high infection, and cluster **3** indicates low fatality and low infection.

Even though we observed a strong correlation between the infection and death rates, this data does not provide information about the disease fatality, that is the proportion of deaths occurring from infections. When the fatality rate (expressed as deaths per 10,000 infections) was calculated for all counties and plotted against the infection rate, a distinct pattern of relationship between these two variables was found (Figure 2B). We observed that the counties with a high fatality rate had a relatively low infection rate while the counties with high infection rates had a relatively low fatality rate. In accordance, when the counties were divided into three clusters on the basis of infection and fatality rate using a two-dimensional k-means clustering method, we obtained clusters with the following features (Figure 2B): (1) High fatality and low infection rate (cluster **1**); (2) high infection and low fatality rate (cluster **2**); (3) low infection and low fatality rate (cluster **3**). Interestingly, the locations of the counties included in cluster **1** (Hamilton, Steuben, Tioga, Yates, Orleans, and Warren) were distributed across the NYS (Figure 2C; counties shaded in red). The counties in cluster **2** (Figure 2C; counties shaded in yellow) were all in proximity and located near the NYC, although in terms of the fatality rate, they were interspersed with cluster **3** (Figure 2C; counties shaded in green). Since high fatality from COVID-19 is observed among nursing residents (Rada, 2020), we also checked whether the distribution of nursing homes has a contribution to the variation of fatality observed between NYS counties. Mapping the nursing homes in individual counties, we did not find an apparent relationship between counties with high fatality and increased density of nursing homes (Figure 2C). These results suggest that various risk factors for COVID-19 have a differential contribution on infection and fatality.

3.2 Impact of demographic factors on COVID-19

Multiple studies have shown the association of demographic variables with COVID-19 infection and outcome (Goldstein and Lee, 2020; Karmakar et al., 2021; Perone, 2021; Sorci et al., 2020). To study how they vary across the clusters of NYS counties that we constructed on the basis of COVID-19 infection, death, and fatality, we selected five well-known demographic risk factors namely, population density, age (percentage of people with age above 55 yr), ethnicity (percentage of African American and Hispanic American population), and poverty (percentage of the population with income below poverty line). Additionally, we considered the distance from the disease epicenter, measured as the distance of a county from Manhattan in NYC. Figure 3 shows these variables plotted against counties organized in three clusters as described in the previous section. Each variable demonstrated a characteristic pattern of distribution within the clusters. KW test followed by multiple comparisons was further performed to calculate the statistical difference.

The trends for most demographic variables followed a similar pattern for infection and death clusters except for poverty. For population density and ethnicity (African American or Hispanic American) median values showed a decreasing trend from cluster **1** to cluster **3**, while an opposite trend was observed for age and distance from the epicenter (Figure 3). Furthermore, the difference between clusters **1** and **2** was not significant for these variables but their difference

with cluster **3** was found to be significant (except between clusters **2** and **3** for age). Interestingly, for the percentage of the population below the poverty line, the highest median value was observed in cluster **2** for the infection group, and in cluster **1** for the death group; the median values in cluster **3** was intermediate for both infection and death groups, thus, suggesting a role of poverty on COVID-19 infection and death cannot be explained through simple association.

The clusters in the fatality group demonstrated a highly distinct pattern of association with demographic variables (Figure 3). For all variables except poverty, the median values of cluster **3** (low fatality and low infection) were found to be intermediate between the median values of cluster **1** (high fatality and low infection) and cluster **2** (low fatality and high infection). Specifically, for the percentage of population with age over 55 yr, the highest median value was observed in cluster **1** (36.5%) in comparison to 28.9% and 33.7% observed in cluster **2** and cluster **3**, respectively. For population density and ethnicity, the trend was opposite with cluster **2** and cluster **1** showing the highest and lowest median values among the clusters, respectively (Figure 3). These findings indicate that high fatality and infections are associated with different sets of demographic risk factors. Overall, the analysis suggests a potential role of demographic structure towards the extent of observed infection, death, and fatality from COVID-19.

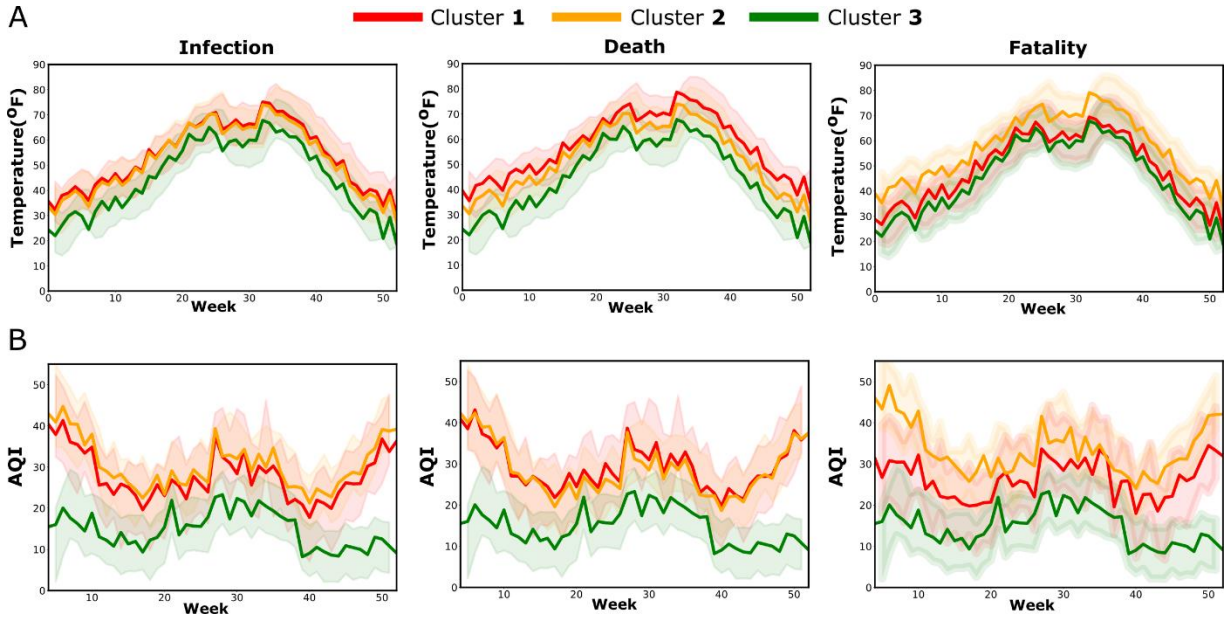


Figure 4. ARIMA time series analysis of temperature (A) and air AQI (B) from weekly EPA data (2015-2019). Predicted values with 95% confidence band for one year, starting from January are shown in the plots. EPA sites for each cluster were located in a representative county belonging to that cluster. For infection and death categories, clusters **1**, **2**, and **3** represent high, intermediate, and low rates of infection or death; for fatality category, cluster **1** indicates high fatality and low infection, cluster **2** indicates low fatality and high infection, and cluster **3** indicates low fatality and low infection. Abbreviations: ARIMA, autoregressive integrated moving average; AQI, air quality index.

3.3 Impact of environmental factors on COVID-19

Since several recent studies have shown an association of environmental factors such as air pollution and temperature on COVID-19 transmission and severity (Li et al., 2020; Wu et al., 2020), we investigated whether such association could be observed across the clusters of NYS counties. Furthermore, we hypothesized chronic exposure to have a stronger impact than acute exposure. This prompted us to select one EPA site representative for each cluster and collect temperature and AQI data for the years 2015-2019. To compare the variables between the sites and find out any differences throughout the year or any specific period of the year, ARIMA models were constructed from weekly time series data. Figure 4 shows ARIMA models of predicted values with 95% confidence bands for temperature and AQI. The AIC values of the models for all conditions were low and comparable (range 289.68 – 303.11), confirming the model robustness. The model predicted temperatures showed a similar pattern for all three clusters in all three categories of infection, death, and fatality with values reaching a peak during the summer months of June-August. Although the predicted temperature for cluster **3** in the infection and the death categories were slightly lower than the other two clusters, there was a considerable overlap of the confidence bands, thus an association the clusters with temperature could not be inferred (Figure 4A). Similarly, the confidence bands of the models for temperature in the fatality clusters also demonstrated a substantial overlap. In contrast to temperature, the model predicted AQI values demonstrated a larger separation between the clusters (Figure 4B). In the infection and death groups, AQI values for cluster **3** were substantially lower than the values for clusters **1** and **2**, and the differences were more prominent during the winter months with separation of the confidence bands. In the fatality group, the highest AQI values were observed in the cluster **2**, which corresponds to high infection but low fatality, and the lowest AQI values were observed in the cluster **3**, corresponding to low infection and low fatality. Thus, the analysis of EPA data suggests COVID-19 in NYS is linked to poor air quality but not with outdoor temperature.

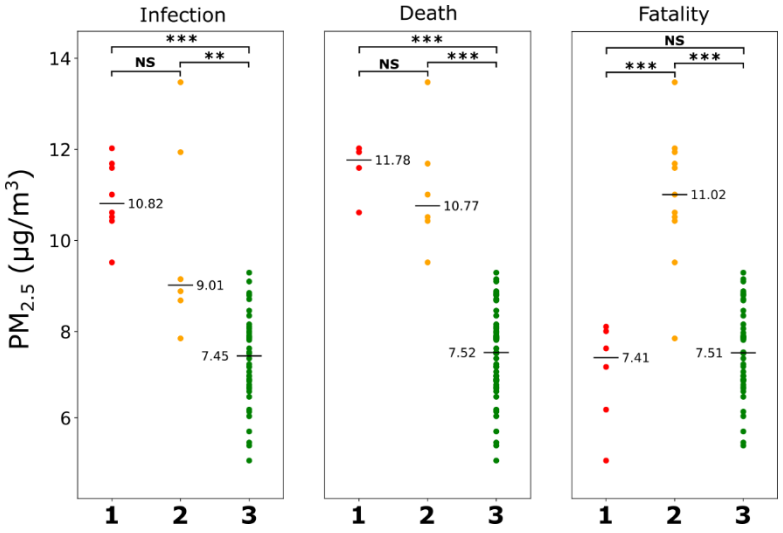


Figure 5. Temporally averaged $PM_{2.5}$ estimates from NYS counties for the years 2000-2016 are compared between clusters based on COVID-19 infection, death, and fatality. Horizontal lines represent median values. ***P < 0.01, **P < 0.05, *P < 0.10, NS not

significant (Mann-Whitney U test with Bonferroni corrections). For infection and death categories, clusters **1**, **2**, and **3** represent high, intermediate, and low rates of infection or death; for fatality category, cluster **1** indicates high fatality and low infection, cluster **2** indicates low fatality and high infection, and cluster **3** indicates low fatality and low infection.

Although EPA measurements provide an accurate estimate of the air quality, data at the resolution of individual counties are not available due to the relatively few EPA sampling sites across the NYS. Therefore, to capture the variation of air quality across the counties, we used temporally averaged $PM_{2.5}$ estimates from satellite data and ground-based measurements over a time period of 2000-2016 (Wu et al., 2020). When the $PM_{2.5}$ values from the counties were compared between the COVID-19 clusters for infection, death, and fatality, the pattern corroborated well with the observations from EPA data (Figure 5). For COVID-19 infection and death, $PM_{2.5}$ values of counties in cluster **3** were significantly lower than the counties in clusters **1** and **2**, with no significant difference observed between the latter two. Similar to the findings with EPA data, in the fatality group, the $PM_{2.5}$ of counties in cluster **2** was significantly higher than in cluster **1** and **3**. These findings demonstrate the association of $PM_{2.5}$ with COVID-19 infection and death in the NYS.

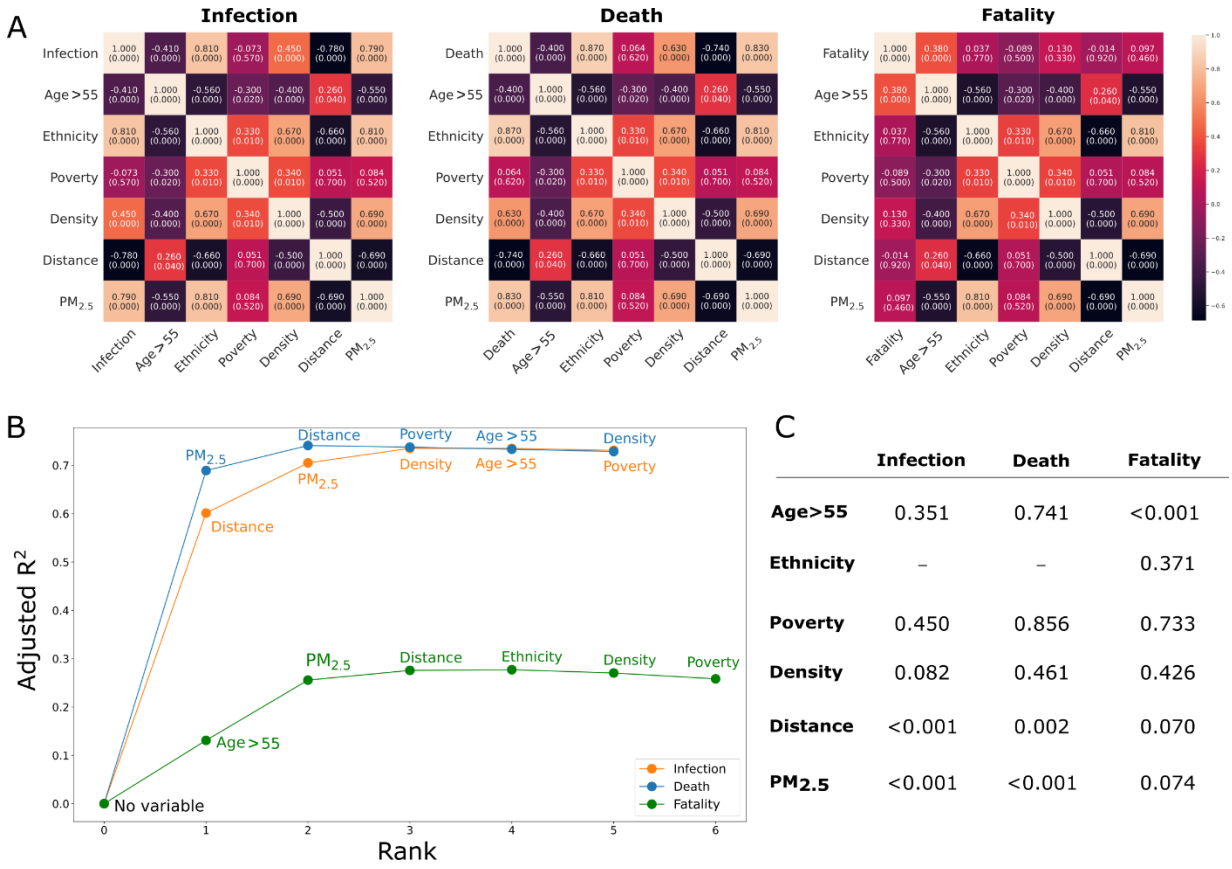
3.4 Contributions of risk factors on COVID-19 infection, death, and fatality

Since clustering of counties based on COVID-19 infection, death, or fatality demonstrated a distinct pattern of association with demographic or environmental risk factors, we wanted to further elucidate the contribution of these variables on the specific aspects of COVID-19 burden. Six risk factors namely, age above 55 yr, ethnicity (African American and Hispanic American population), poverty, population density, distance from the epicenter, and $PM_{2.5}$ were considered as predictor variables, and multivariate regression model with forward "stepwise" selection was used for analysis. Three separate models were constructed using infection, death, and fatality rate as dependent variables to understand the relative contribution of the risk factors for each of these outcomes. Rockland county was excluded from the models as it was identified as an outlier while performing residual analyses of the regression output.

Multicollinearity among the predictor variables can lead to unstable and unreliable estimates of regression coefficients, reducing the power of the regression model. Therefore, before their incorporation in the regression models, we checked for multicollinearity. The correlation matrices in Figure 6A show the Pearson's correlations coefficients among variable pairs. A strong positive correlation was found between ethnicity and $PM_{2.5}$ ($r = 0.81$, $P < 0.001$), while moderate positive correlations were observed between population density and $PM_{2.5}$, ($r = 0.69$, $P < 0.001$) or ethnicity ($r = 0.67$, $P < 0.001$). Additionally, ethnicity and $PM_{2.5}$ demonstrated a strong positive correlation with infection and death, while the distance from the epicenter held a strong negative correlation with these dependent variables. Interestingly, such strong correlations were not observed for fatality, the third dependent variable.

The existence of multicollinearity among predictor variables prompted us to calculate the VIF for each variable to assess their suitability for inclusion in the regression model. We found that VIFs of all variables were lower than the acceptable cut-off value of 5, except for ethnicity when

infection or death was used as dependent variables. This implies that ethnicity is not an independent predictor for infection and death, and therefore, was excluded in the regression models for these two variables. The models revealed distinct contributions of the predictor variables to infection, death, and fatality rates (Figure 6B). $PM_{2.5}$ and distance from the epicenter were found to be the two most important predictors for infection and death. For infection rate, distance from the epicenter was the strongest predictor with a highly significant regression coefficient ($P<0.001$, Figure 6C) and generated an adjusted R^2 value of 0.60 when considered as a sole contributor of the model (Figure 6B); the adjusted R^2 value increased to 0.71 following the inclusion of $PM_{2.5}$ in the model, which also had a significant regression coefficient ($P<0.001$; Figure 6B, C). Regression coefficients of population density, age, and poverty emerged as not significant ($P>0.05$) and their addition to the regression model only marginally increased the adjusted R^2 value 0.74. Similar to infection, distance from the epicenter and $PM_{2.5}$ were two major predictors for the death rate, however, $PM_{2.5}$ was the strongest among them contributing to an adjusted R^2 value of 0.69. The value increased to 0.73 following the inclusion of distance from the epicenter in the model but did not change further upon the addition of other variables.



C

	Infection	Death	Fatality
Age>55	0.351	0.741	<0.001
Ethnicity	–	–	0.371
Poverty	0.450	0.856	0.733
Density	0.082	0.461	0.426
Distance	<0.001	0.002	0.070
$PM_{2.5}$	<0.001	<0.001	0.074

Figure 6. Regression analysis to assess relative contribution of risk factors on COVID-19 infection, death, and fatality. (A) Correlation matrices between demographic risk variables and $PM_{2.5}$ for infection, death, and fatality. Pearson's correlations coefficients between the variables are shown with corresponding P values are mentioned in the parentheses. (B) Stepwise regression model with forward selection for infection, death, and fatality. (Note that

ethnicity is excluded from the models for infection and death due to the existence of strong multicollinearity). (C) Table showing the P values of the regression coefficients from the analysis. Abbreviations used: Density, population density (per mi^2); Ethnicity, African American and Hispanic American (%); Poverty, population below poverty line (%); Distance, distance from the epicenter (mi).

Unlike the models for infection and death rates, age over 55 yr was found to be the strongest predictor in the model for fatality rate, (Figure 6B, C). $PM_{2.5}$ was the second major predictor in the model, although the relatively high P-value (0.074) of the regression coefficient suggests a weaker link with fatality. It should also be noted that the adjusted R^2 value of the final model for fatality was 0.26, indicating that the goodness of model fit is much lower than the other two models for infection and death. Together, the regression analysis helped to delineate the risk factors with major contributions on specific aspects of the COVID-19 burden.

4 Discussion

Our analysis has shown a wide heterogeneity in the infection, death, and fatality rates from COVID-19 among the counties in the NYS during the first pandemic wave. Infection was found to be strongly correlated with death but not with fatality. By grouping the counties into clusters, we show the association of multiple demographic factors and air quality with infection, death, and fatality from COVID-19. Specifically, $PM_{2.5}$, population density, and proportion of African American or Hispanic American population demonstrated a positive association with infection and death, while the distance from the disease epicenter showed a negative association. In contrast, higher fatality from the disease was primarily associated with a higher proportion of the population aged above 55 yr. Furthermore, regression analysis has identified the major contributors among these risk factors for infection, death, and fatality. These results could help to better understand the impact of environmental and demographic factors on COVID-19 in NYS.

Our analysis shows an association of $PM_{2.5}$ with infection and death, a potential but weaker link to fatality. This finding is in agreement with studies focused on the role of outdoor air pollution on COVID-19 transmission and health outcomes (Benmarhnia, 2020; Gupta et al., 2020; Lolli et al., 2020; Pozzer et al., 2020; Wu et al., 2020). $PM_{2.5}$ has been reported to be positively correlated with both increased COVID-19 transmission and fatality. Comorbid conditions such as respiratory and cardiovascular illnesses that are associated with chronic exposure to higher $PM_{2.5}$ are thought to aggravate the illness from virus infection, posing a higher risk of death (Benmarhnia, 2020; Pozzer et al., 2020; Wu et al., 2020). Additional mechanisms proposed for increased SARS-CoV-2 transmission by $PM_{2.5}$ include the particulate matters serving as a transport vector and their ability to increase the susceptibility to infection by inducing lung inflammation (Maleki et al., 2021). It is to be noted that although the average $PM_{2.5}$ values in the NYS meet the safe limit (< 12 $\mu g/m^3$) set by EPA (Jin et al., 2019), an association of $PM_{2.5}$ level with adverse COVID-19 outcome can be clearly discerned. We attribute this apparent discrepancy to the existence of potential hot spots where the exposure $PM_{2.5}$ could be much higher than the average. Although our analysis considers chronic exposure to $PM_{2.5}$ (average values from 2000-2016) to better capture the long-term health effects(Wu et al., 2020), a recent

study showed no significant difference in air quality in the NYC area after lockdown (Zangari et al., 2020), and thus would also reasonably reflect the exposure to $PM_{2.5}$ during the period of investigation. Also, our work focuses on $PM_{2.5}$ alone, however, other air pollutants, especially NO_2 and O_3 are reported to be associated with higher COVID-19 spread and fatality (Adhikari and Yin, 2020; Copat et al., 2020; Liang et al., 2020). Commonly generated by anthropogenic activities such as traffic, NO_2 is a well-known inducer of lung inflammation and can synergistically act with $PM_{2.5}$ to increase the adverse impact of COVID-19 (Hesterberg et al., 2009; Huang et al., 2012). NO_2 is also a source for O_3 formation, which is further facilitated by higher temperature and $PM_{2.5}$ (Zhang et al., 2019). O_3 is a strong oxidant and exposure to O_3 is known to cause or aggravate respiratory and cardiovascular diseases, and older adults are shown to be more vulnerable to the adverse effects (Day et al., 2018; Zhang et al., 2019). Thus, apart from independent effects, the interactions of NO_2 and O_3 with $PM_{2.5}$ could be important in the context of COVID-19 and needs to be investigated in the future study. Outdoor air temperature, the other environmental parameter we examined in this study did not show any association with COVID-19 cases; both positive and negative association of temperature with COVID-19 infection have been reported in the literature, and a recent systematic review concluded data-related and methodological issues including inherent uncertainties of the data, inappropriate controlling for confounding parameters, and short periods of investigation underlie such conflicting results (Dong et al., 2021)(Lolli et al., 2020). Thus, to better understand the influence of temperature on COVID-19, future studies should be conducted with time-resolved data for a longer period and taking appropriate measures for confounding variables.

We observed a strong correlation between $PM_{2.5}$ and the percentage of the population belonging to African American or Hispanic American ethnicity ($r = 0.81$, $P < 0.001$), suggesting that people from these ethnicities are exposed to a higher level of $PM_{2.5}$ than the average population. Multiple studies have concluded that these two ethnicities in the US are at disproportionately higher risk of COVID-19 (Cordes and Castro, 2020; Li et al., 2020; Martinez et al., 2020; Yancy, 2020). Factors related to socioeconomic inequities such as the greater risk of virus exposure from professional demand or living in crowded accommodation, higher prevalence of chronic comorbidity, and restricted access to healthcare are thought to underlie such differences (Patel et al., 2020). Our results suggest that exposure to air pollution could be a contributor to further increase this disparity. Low socioeconomic status is thought to pose a greater risk for COVID-19 exposure (Yancy, 2020); however, for NYS, we observed counties in cluster **3** for infection and death rates to have a higher proportion of people living below the poverty line than the counties from other two clusters. That cluster **3** counties have a relatively lower risk from other demographic and environmental variables could explain this apparent discrepancy. Indeed, when considering the population of NYC alone, poverty and COVID-19 are found to be positively correlated (Cordes and Castro, 2020).

Two demographic variables, distance from the epicenter and age above 55 years, came out as major contributors in our regression models. However, their influences on COVID-19 were distinct. The distance of counties from the disease epicenter was inversely related to infection and death, alone accounting for 60% of variation in the regression model of infection. This finding is not unusual as the disease spread would be facilitated by the population mobility with the highest effect on the neighboring regions. We also observed a strong correlation between

infection and death rates ($r = 0.92$, $P < 0.0001$), corroborating their association with a similar set of risk factors. In contrast, a poor correlation was observed between infection and fatality, and the regression model for fatality revealed age over 55 years to be the most significant independent variable. Fatality from COVID-19 depends on the health of patients where age plays a crucial role (Mesas et al., 2020; Richardson et al., 2020). Increased risk of the aged population to complications and death from COVID-19 is observed across the world, and multiple factors including the existence of chronic comorbid conditions and a weaker immune system are thought to underlie such vulnerability (Mesas et al., 2020). The association of distinct sets of risk factors for death and fatality suggests that they should be considered as separate metrics for the COVID-19 burden for the development of preventive or mitigative strategies.

Grouping the counties into clusters not only helped to visualize how NYS counties are impacted by specific COVID-19 adversities but also allowed easier comparison of their association with various demographic and environmental risk variables. While the regression models have further helped to identify the risk variables that have major contributions to specific aspects of disease impact, it should also be noted that the adjusted R^2 value of the regression model for fatality is substantially lower than the models for infection and death (0.26 *vs.* 0.74 and 0.73). This difference suggests the possibility of missing key variables in the model for COVID-19 fatality that needs to be identified and incorporated in the future model. Such variables could potentially be measures pertaining to the outcome of an infected individual, including the availability and access to healthcare resources, vaccination, and awareness for early diagnosis and treatment.

Conclusions

In this work, we analyzed the association of multiple demographic and environmental factors with the COVID-19 burden in NYS during the first pandemic wave. Clustering the counties based on COVID-19 infection or death revealed their segregation by geographical location with clusters located farther away from NYC showing lower infection or death. In contrast, counties grouped in the cluster for high disease fatality were distributed across the NYS and were different from those having high infection and death rates. The clustered counties showed a prominent association with demographic variables and $PM_{2.5}$ but the patterns of association for infection and death were distinct than for fatality. Clusters with high infection and death were found to have higher $PM_{2.5}$, higher population density, a higher proportion of African Americans and Hispanic Americans, and were closer to the disease epicenter, while the cluster with higher fatality had a higher proportion of population aged above 55 yr. Stepwise regression models built on county data further showed that $PM_{2.5}$ and the distance from the epicenter are two major contributors for infection and death, while advanced age makes the strongest contribution to fatality. Although our study is confined to counties within the NYS, we observed prominent differences in the distribution of infection and fatality along with an association with distinct sets of demographic and environmental risk variables. The US being a country with a vast size, have considerable heterogeneity between states in terms of social and cultural practices, public health policies, access to healthcare, and general awareness of COVID-19, all of which could have a significant impact on the absolute magnitude of COVID-19 burden; however, we expect that the variables considered in this work would still have similar effects as observed for the NYS, and

thus, our results can provide key insight on the contribution of demographic and environmental factors on the disease landscape in these states. Additionally, a similar modeling approach could be utilized in future studies to include additional relevant variables in the analysis to understand their contribution to the disease. With strong anthropogenic contributions to the environment in modern societies, our findings suggest the need for critical consideration of both demographic and environmental variables when predicting the impact of COVID-19 or developing preventive or mitigative strategies to control the disease.

Funding information

Not applicable.

Declaration of conflicts of interest

The authors declared that they have no conflicts of interest.

Acknowledgments

Vijay Kumar acknowledges the support from US-Pakistan Knowledge Corridor PhD Scholarship Program under Higher Education Commission, Pakistan. Bridget Wangler thanks the Clarkson University Honors Program for their support.

References

Adhikari A, Yin J. Short-Term Effects of Ambient Ozone, PM2.5, and Meteorological Factors on COVID-19 Confirmed Cases and Deaths in Queens, New York. Int J Environ Res Public Health 2020; 17: 4047.

Arif M, Sengupta S. Nexus between population density and novel coronavirus (COVID-19) pandemic in the south Indian states: A geo-statistical approach. Environment, Development and Sustainability 2020: 1-29.

Auger KA, Shah SS, Richardson T, Hartley D, Hall M, Warniment A, et al. Association Between Statewide School Closure and COVID-19 Incidence and Mortality in the US. JAMA 2020; 324: 859-870.

Baldwin R, Di Mauro BW. Economics in the time of COVID-19: A new eBook: CEPR Press, 2020.

Bashir MF, Ma BJ, Bilal, Komal B, Bashir MA, Tan DJ, et al. Correlation between climate indicators and COVID-19 pandemic in New York, USA. Science of the Total Environment 2020; 728: 138835.

Benmarhnia T. Linkages Between Air Pollution and the Health Burden From COVID-19: Methodological Challenges and Opportunities. American Journal of Epidemiology 2020; 189: kwaa148.

Brook RD, Rajagopalan S, Pope CA, Brook JR, Bhatnagar A, Diez-Roux AV, et al. Particulate Matter Air Pollution and Cardiovascular Disease. Circulation 2010; 121: 2331-2378.

Chatterjee S, Simonoff JS. Handbook of regression analysis. Vol 5: John Wiley & Sons, 2013.

Chauhan AJ, Johnston SL. Air pollution and infection in respiratory illness. Br Med Bull 2003; 68: 95-112.

Chen JT, Krieger N. Revealing the Unequal Burden of COVID-19 by Income, Race/Ethnicity, and Household Crowding: US County Versus Zip Code Analyses. J Public Health Manag Pract 2021; 27 Suppl 1, COVID-19 and Public Health: Looking Back, Moving Forward: S43-S56.

Copat C, Cristaldi A, Fiore M, Grasso A, Zuccarello P, Signorelli SS, et al. The role of air pollution (PM and NO2) in COVID-19 spread and lethality: A systematic review. Environ Res 2020; 191: 110129.

Copiello S, Grillenzoni C. The spread of 2019-nCoV in China was primarily driven by population density. Comment on "Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China" by Zhu et al. Science of The Total Environment 2020; 744: 141028.

Cordes J, Castro MC. Spatial analysis of COVID-19 clusters and contextual factors in New York City. Spatial and Spatio-temporal Epidemiology 2020; 34: 100355.

Day DB, Clyde MA, Xiang J, Li F, Cui X, Mo J, et al. Age modification of ozone associations with cardiovascular disease risk in adults: a potential role for soluble P-selectin and blood pressure. J Thorac Dis 2018; 10: 4643-4652.

Dong ZM, Fan XR, Wang J, Mao YX, Luo YY, Tang S. Data-related and methodological obstacles to determining associations between temperature and COVID-19 transmission. Environmental Research Letters 2021; 16: 034016.

Donkelaar Av, Martin RV, Li C, Burnett RT. Regional Estimates of Chemical Composition of Fine Particulate Matter Using a Combined Geoscience-Statistical Method with Information from Satellites, Models, and Monitors. Environmental Science & Technology 2019; 53: 2595-2611.

Fahim AM, Salem AM, Torkey FA, Ramadan MA. An efficient enhanced k-means clustering algorithm. Journal of Zhejiang University-SCIENCE A 2006; 7: 1626-1633.

Feng C, Li J, Sun W, Zhang Y, Wang Q. Impact of ambient fine particulate matter (PM2.5) exposure on the risk of influenza-like-illness: a time-series analysis in Beijing, China. Environ Health 2016a; 15: 17.

Feng S, Gao D, Liao F, Zhou F, Wang X. The health effects of ambient PM2.5 and potential mechanisms. Ecotoxicol Environ Saf 2016b; 128: 67-74.

Goldstein JR, Lee RD. Demographic perspectives on the mortality of COVID-19 and other epidemics. Proc Natl Acad Sci U S A 2020; 117: 22035-22041.

Guan WJ, Zheng XY, Chung KF, Zhong NS. Impact of air pollution on the burden of chronic respiratory diseases in China: time for urgent action. Lancet 2016; 388: 1939-1951.

606 Gupta A, Bherwani H, Gautam S, Anjum S, Musugu K, Kumar N, et al. Air pollution
607 aggravating COVID-19 lethality? Exploration in Asian cities using statistical models.
608 Environment, Development and Sustainability 2020: 1-10.
609 Hesterberg TW, Bunn WB, McClellan RO, Hamade AK, Long CM, Valberg PA. Critical review
610 of the human data on short-term nitrogen dioxide (NO2) exposures: evidence for NO2
611 no-effect levels. Crit Rev Toxicol 2009; 39: 743-81.
612 Hopke PK, Croft D, Zhang W, Lin S, Masiol M, Squizzato S, et al. Changes in the acute
613 response of respiratory diseases to PM2.5 in New York State from 2005 to 2016. Sci
614 Total Environ 2019; 677: 328-339.
615 Huang YC, Rappold AG, Graff DW, Ghio AJ, Devlin RB. Synergistic effects of exposure to
616 concentrated ambient fine pollution particles and nitrogen dioxide in humans. Inhal
617 Toxicol 2012; 24: 790-7.
618 Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice: OTexts, 2018.
619 Jin XM, Fiore AM, Civerolo K, Bi JZ, Liu Y, van Donkelaar A, et al. Comparison of multiple
620 PM2.5 exposure products for estimating health benefits of emission controls over New
621 York State, USA. Environmental Research Letters 2019; 14: 084023.
622 Karmakar M, Lantz PM, Tipirneni R. Association of Social and Demographic Factors With
623 COVID-19 Incidence and Death Rates in the US. JAMA Netw Open 2021; 4: e2036462.
624 Lee VJ, Chiew CJ, Khong WX. Interrupting transmission of COVID-19: lessons from
625 containment efforts in Singapore. J Travel Med 2020; 27.
626 Li AY, Hannah TC, Durbin JR, Dreher N, McAuley FM, Marayati NF, et al. Multivariate
627 Analysis of Black Race and Environmental Temperature on COVID-19 in the US. Am J
628 Med Sci 2020; 360: 348-356.
629 Liang D, Shi L, Zhao J, Liu P, Sarnat JA, Gao S, et al. Urban Air Pollution May Enhance
630 COVID-19 Case-Fatality and Mortality Rates in the United States. The Innovation 2020;
631 1: 100047.
632 Liu J, Zhou J, Yao J, Zhang X, Li L, Xu X, et al. Impact of meteorological factors on the
633 COVID-19 transmission: A multi-city study in China. Science of The Total Environment
634 2020; 726: 138513.
635 Lolli S, Chen YC, Wang SH, Vivone G. Impact of meteorological conditions and air pollution
636 on COVID-19 pandemic transmission in Italy. Sci Rep 2020; 10: 16213.
637 Lusignan Sd, Dorward J, Correa A, Jones N, Akinyemi O, Amirthalingam G, et al. Risk factors
638 for SARS-CoV-2 among patients in the Oxford Royal College of General Practitioners
639 Research and Surveillance Centre primary care network: a cross-sectional study. The
640 Lancet Infectious Diseases 2020; 20: 1034-1042.
641 Maleki M, Anvari E, Hopke PK, Noorimotlagh Z, Mirzaee SA. An updated systematic review on
642 the association between atmospheric particulate matter pollution and prevalence of
643 SARS-CoV-2. Environmental Research 2021; 195: 110898.

644 Martinez DA, Hinson JS, Klein EY, Irvin NA, Saheed M, Page KR, et al. SARS-CoV-2
645 Positivity Rate for Latinos in the Baltimore-Washington, DC Region. JAMA 2020; 324:
646 392-395.
647 Mesas AE, Cavero-Redondo I, Alvarez-Bueno C, Sarria Cabrera MA, Maffei de Andrade S,
648 Sequi-Dominguez I, et al. Predictors of in-hospital COVID-19 mortality: A
649 comprehensive systematic review and meta-analysis exploring differences by age, sex
650 and health conditions. PLoS One 2020; 15: e0241742.
651 Miller LE, Bhattacharyya R, Miller AL. Data regarding country-specific variability in Covid-19
652 prevalence, incidence, and case fatality rate. Data in Brief 2020; 32: 106276.
653 Monmonier M, Giordano A. GIS in New York State county emergency management offices:
654 User assessment. Applied Geographic Studies 1998; 2: 95-109.
655 Patel JA, Nielsen FBH, Badiani AA, Assi S, Unadkat VA, Patel B, et al. Poverty, inequality and
656 COVID-19: the forgotten vulnerable. Public Health 2020; 183: 110-111.
657 Perone G. The determinants of COVID-19 case fatality rate (CFR) in the Italian regions and
658 provinces: An analysis of environmental, demographic, and healthcare factors. Science of
659 The Total Environment 2021; 755: 142523.
660 Pozzer A, Dominici F, Haines A, Witt C, Munzel T, Lelieveld J. Regional and global
661 contributions of air pollution to risk of death from COVID-19. Cardiovasc Res 2020;
662 116: 2247-2253.
663 Pradhan A, Olsson PE. Sex differences in severity and mortality from COVID-19: are males
664 more vulnerable? Biol Sex Differ 2020; 11: 53.
665 Rada AG. Covid-19: the precarious position of Spain's nursing homes. BMJ 2020; 369: m1554.
666 Reichberg SB, Mitra PP, Haghamad A, Ramrattan G, Crawford JM, Northwell C-RC, et al.
667 Rapid Emergence of SARS-CoV-2 in the Greater New York Metropolitan Area:
668 Geolocation, Demographics, Positivity Rates, and Hospitalization for 46 793 Persons
669 Tested by Northwell Health. Clin Infect Dis 2020; 71: 3204-3213.
670 Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, et al.
671 Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients
672 Hospitalized With COVID-19 in the New York City Area. JAMA 2020; 323: 2052-2059.
673 Rocklöv J, Sjödin H. High population densities catalyze the spread of COVID-19. Journal of
674 Travel Medicine 2020; 27.
675 Sarkodie SA, Owusu PA. Global assessment of environment, health and economic impact of the
676 novel coronavirus (COVID-19). Environ Dev Sustain 2020a: 1-11.
677 Sarkodie SA, Owusu PA. Impact of meteorological factors on COVID-19 pandemic: Evidence
678 from top 20 countries with confirmed cases. Environ Res 2020b; 191: 110101.
679 Sorci G, Faivre B, Morand S. Explaining among-country variation in COVID-19 case fatality
680 rate. Sci Rep 2020; 10: 18909.

U.S. Census Bureau ACS. American Community Survey 1-Year Estimates, Table DP05. Retrieved from <https://data.census.gov/cedsci/table?q=dp05&tid=ACSDP1Y2018.DP05>, 2018.

Wadhera RK, Wadhera P, Gaba P, Figueroa JF, Joynt Maddox KE, Yeh RW, et al. Variation in COVID-19 Hospitalizations and Deaths Across New York City Boroughs. JAMA 2020; 323: 2192-2195.

Wellenius GA, Burger MR, Coull BA, Schwartz J, Suh HH, Koutrakis P, et al. Ambient air pollution and the risk of acute ischemic stroke. Arch Intern Med 2012; 172: 229-34.

Wu X, Nethery RC, Sabath MB, Braun D, Dominici F. Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis. Sci Adv 2020; 6: eabd4049.

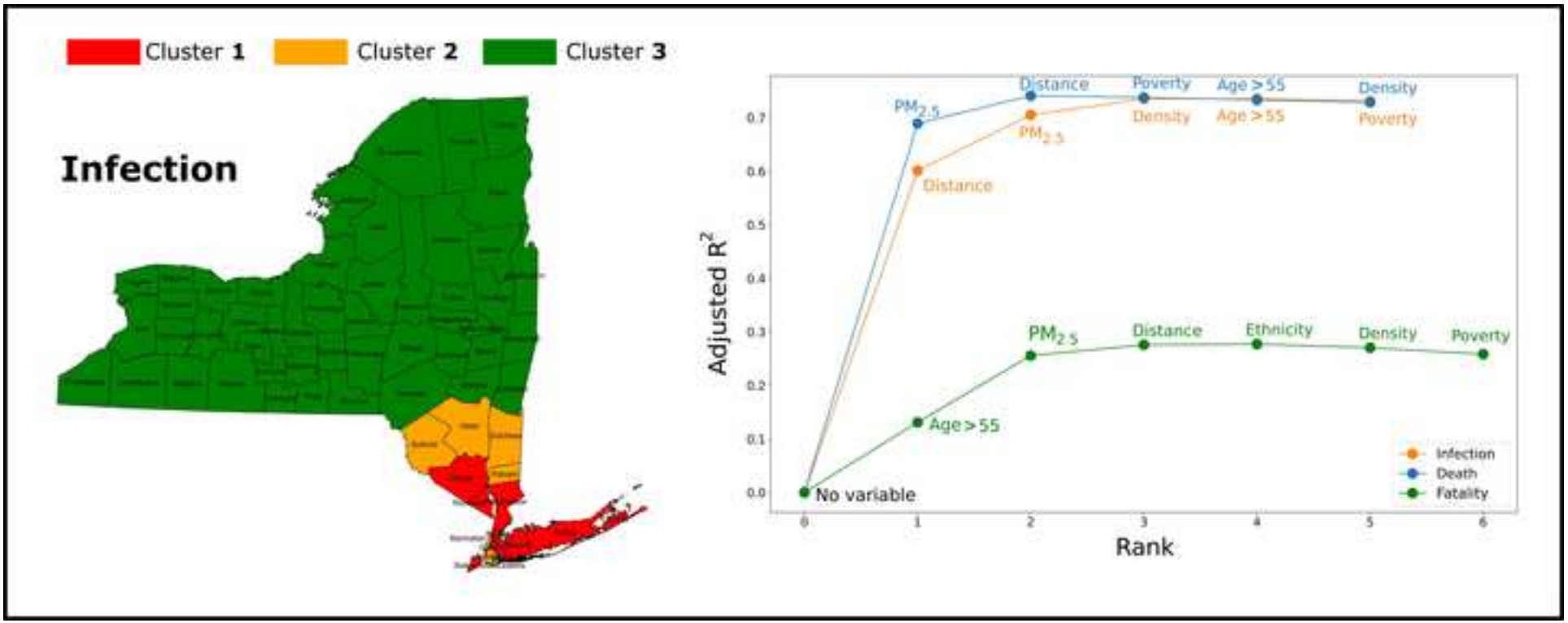
Xing YF, Xu YH, Shi MH, Lian YX. The impact of PM2.5 on the human respiratory system. J Thorac Dis 2016; 8: E69-74.

Yancy CW. COVID-19 and African Americans. JAMA 2020; 323: 1891-1892.

Zangari S, Hill DT, Charette AT, Mirowsky JE. Air quality changes in New York City during the COVID-19 pandemic. Sci Total Environ 2020; 742: 140496.

Zhang JJ, Wei Y, Fang Z. Ozone Pollution: A Major Health Hazard Worldwide. Front Immunol 2019; 10: 2518.

Graphical Abstract



Highlights (for review : 3 to 5 bullet points (maximum 85 characters including spaces per bullet point)

Highlights:

- The impact of COVID-19 varied widely across the counties in the state of New York during the first pandemic wave.
- Long-term exposure to higher $PM_{2.5}$ is associated an increased infection, mortality, and fatality.
- Distance from the disease epicenter strongly influences COVID-19 infection burden.
- Increased age (>55 year) is found to be the strongest predictor for COVID-19 fatality.

ABSTRACT

The coronavirus disease 2019 (COVID-19) has had a global impact that has been unevenly distributed amongst and, even within countries. Multiple demographic and environmental factors have been associated with the risk of COVID-19 spread and fatality, including age, gender, ethnicity, poverty, and air quality among others. However, specific contributions of these factors are yet to be understood. Here, we attempted to explain the variability in infection, death, and fatality rates by understanding the contributions of a few selected factors. We compared the incidence of COVID-19 in New York State (NYS) counties during the first wave of infection and analyzed how different demographic and environmental variables associate with the variation observed across the counties. We observed that infection and death rates, two important COVID-19 metrics, to be highly correlated with both being highest in counties located near New York City, considered as one of the epicenters of the infection in the US. In contrast, disease fatality was found to be highest in a different set of counties despite registering a low infection rate. To investigate this apparent discrepancy, we divided the counties into three clusters based on COVID-19 infection, death rate, or fatality, and compared the differences in the demographic and environmental variables such as ethnicity, age, population density, poverty, temperature, and air quality in each of these clusters. Furthermore, a regression model built on this data reveals $PM_{2.5}$ and distance from the epicenter are significant risk factors for infection, while disease fatality has a strong association with age and $PM_{2.5}$. Our results demonstrate that for the NYS, demographic components distinctly associate with specific aspects of COVID-19 burden and also highlight the detrimental impact of poor air quality. These results could help design and direct location-specific control and mitigation strategies.

Keywords: COVID-19, New York State, air quality, $PM_{2.5}$, clustering, stepwise regression

1 Introduction

The impact of the COVID-19 pandemic on global health and economy has exceeded well over the severity of any other communicable diseases in recent history (Baldwin and Di Mauro, 2020; Sarkodie and Owusu, 2020a). The pandemic has also stimulated and significantly accelerated global research into coronaviruses, airborne disease transmissions, and development of new vaccines. Within a short span of time, scientists have succeeded in obtaining critical information on the structure and genomic sequence of the virus pathogen SARS-CoV-2, mechanism of virus infection to host, modes of transmission, and injury to host organs induced by the virus. The research findings have accelerated the development of vaccines and established preventive measures such as the use of masks. Simultaneously, there has been a significant effort to understand the association of COVID-19 to demographic and environmental factors, to explain the geographical or seasonal variability in disease burden (Goldstein and Lee, 2020; Karmakar et al., 2021; Perone, 2021; Sorci et al., 2020). Underscoring precise influences of human demographics and environmental factors on the pandemic would be important toward developing effective public health and social measures.

Among the demographic variables, age, gender, ethnicity, and population density are reported to impact COVID-19. Advanced age is shown to significantly increase the fatality from COVID-19. A study conducted on hospitalized patients in the New York City (NYC) area found 84% of the total deaths occurred in people aged above 60 years (Mesas et al., 2020; Richardson et al., 2020). Moreover, males were seen to be more susceptible to suffer from COVID-19 complications and fatality (Pradhan and Olsson, 2020). Although the mechanism underlying such predisposition of age and sex is not completely understood, the presence of preexisting health conditions and a lowered immunity associated with higher age are thought to be two major factors (Mesas et al., 2020; Pradhan and Olsson, 2020; Richardson et al., 2020). Chronic comorbidities such as hypertension, ischemic heart disease, diabetes, and chronic obstructive pulmonary disease (COPD) are more common in older age and poses risk for severe outcomes (Lusignan et al., 2020; Richardson et al., 2020). Studies focused on the impact of COVID-19 on the ethnic composition also revealed vulnerabilities of certain ethnicities to the disease. In the US, a disproportionately higher number of COVID-19 infections and deaths are observed among African Americans and Hispanic Americans relative to their share of population (Martinez et al., 2020; Yancy, 2020). Socioeconomic disparities leading to increased exposure and lower access to healthcare are thought to contribute to such vulnerability. High population density is reported to increase the risk of COVID-19 spread (Arif and Sengupta, 2020; Copiello and Grillenzoni, 2020), although it is not the sole determining factor as many dense metropolitan cities in Japan, South Korea, China, and Singapore have observed a low infection rate (Lee et al., 2020; Rocklöv and Sjödin, 2020).

The association of environmental factors such as air quality and meteorological parameters to the adverse effects of COVID-19 has been investigated in multiple studies. Air pollution is of particular interest as chronic exposure to air pollutants is linked to multiple chronic respiratory and cardiovascular diseases such as COPD, ischemic heart disease, and hypertension—diseases which are known to increase COVID-19 fatality (Feng et al., 2016b; Guan et al., 2016; Wellenius et al., 2012). Additionally, air pollution substantially increases the risk of respiratory infections including viral infections (Chauhan and Johnston, 2003; Feng et al., 2016a). Fine particulate matter in the air, especially $PM_{2.5}$ (particulate matter with aerodynamic diameter 2.5 µm or less) has been linked to many of these pollution-mediated health effects (Brook et al., 2010; Hopke et al., 2019; Xing et al., 2016). Early reports indicate a positive association of $PM_{2.5}$ with both COVID-19 transmission and fatality (Gupta et al., 2020; Lolli et al., 2020; Pozzer et al., 2020; Wu et al., 2020). Analysis of meteorological factors based on the data from 30 Chinese cities revealed low temperature, less diurnal temperature variation, and low humidity favor the transmission of COVID-19 infection (Liu et al., 2020). This finding was supported by a larger-scale study using data from the top 20 countries with infections, and further claimed low wind speed, surface pressure, and precipitation to increase the risk of disease spread (Sarkodie and Owusu, 2020b).

While the connection of COVID-19 with demographic and environmental factors has been demonstrated by multiple studies, majority of these studies focused on disease transmission or disease fatality alone, and the analyses were directed to either the demographic or the environmental variables. Since anthropogenic factors have a substantial impact on the environmental variables, consideration of both demographic and environmental factors in the

analysis is expected to increase the robustness of inference and reduce the risk of any spurious association (Copiello and Grillenzoni, 2020). Collating information from existing studies to understand the relative impact of these risk factors on infection burden and disease fatality is challenging since these studies were conducted in different geographical locations, and multiple additional confounding factors such as testing and screening strategies, healthcare infrastructure, and socio-cultural practices could contribute to the wide variability of COVID-19 infection and fatality observed across countries or even between different regions within a country (Auger et al., 2020; Chen and Krieger, 2021; Miller et al., 2020). Therefore, to assess the influence of both demographic and environmental factors on COVID-19, ideally the data should be from a geographical location where these factors show considerable variation with low variability of other confounding factors. New York State (NYS), located in the USA fits well as a potential location to conduct such study as it offers a wide range of variation in its demographic landscapes with urban, population-dense, ethnically diverse counties near NYC to many rural, white-dominated, population-sparse counties located in the upstate region. The $PM_{2.5}$ distribution across the state demonstrates a consistent pattern with distinct variation across regions (Jin et al., 2019). Additionally, state-wide implemented policies, including public health measures, and hospital care would help to reduce the potential differences due to the confounding factors mentioned above. Analyses conducted in the NYC area revealed that multiple demographic factors such as ethnicity, male gender, poverty, and household crowding are associated with increased COVID-19 infection, hospitalization, or death (Chen and Krieger, 2021; Reichberg et al., 2020). Studies from NYC metropolitan area also suggested a potential connection of air quality and meteorological variables such as temperature and humidity to higher COVID-19 transmission (Adhikari and Yin, 2020; Bashir et al., 2020). Inclusion of data from the entire NYS is expected to capture a wider variability of these variables and provide a deeper insight into their role in the COVID-19 burden.

In this work, we considered the NYS data at county-level resolution and attempted to relate the variability in infection, death, and fatality with selected demographic and environmental factors during the first COVID-19 wave. Using publicly available data, we first grouped the counties into clusters based on COVID-19 infection, death, or fatality rates during the study period, and investigated the association of these clusters with various demographic factors (e.g., population density, the proportion of African American and Hispanic American population) and environmental factors (e.g., $PM_{2.5}$ and temperature). To identify the risk variables that have major contributions on specific aspects of COVID-19 burden, regression models were then built using data from individual counties, where infection, death, or fatality rates were considered as response variables while demographic and environmental factors were used as predictor variables.

2 Methods

2.1 Study Area, Data Source, and Variables

For this study, COVID-19 infection and death count during the period of March 1, 2020, to May 16, 2020, were obtained for all 62 counties in the NYS from publicly accessible information available at Syracuse.com. The population estimates for each county were obtained from the

2018 US Census Bureau's American Community Survey (ACS) website (https://www.census.gov/programs-surveys/acs)(U.S. Census Bureau, 2018). Infection and death rates from COVID-19 for each county were calculated by dividing the cumulative infection and cumulative death counts during the study period by the total population of the county, and expressed as number per 100,000 population. The fatality rate of a county was obtained by dividing the cumulative death count by cumulative infection count during the study period and presented as the number of deaths per 10,000 infected population. In addition to the total population, the ACS census database was used to collect the following information for each county:(1) Area; (2) population with age ≥ 55 years; (3) poverty levels; (4) Hispanic American population (Martinez et al., 2020); and (5) African American population (Yancy, 2020). From this information (1) population density (population/square mile), (2) proportion of the population with ≥ 55 years (expressed as %), (3) proportion of Hispanic American (expressed as %), and (4) proportion of African American (expressed as %) population was calculated for each county. All factors except population density and distance from the epicenter were converted to percentages by county. The nursing home locations across the NYS counties were obtained from the Department of Health and Human Services. The data was retrieved through ArcGIS Map 10.7.1 (Monmonier and Giordano, 1998). The distance of a county from Manhattan, located at the center of NYC (considered as the disease epicenter (Reichberg et al., 2020; Wadhera et al., 2020)), was used as the distance of the county from the disease epicenter and was calculated by measuring the distance between the centroids of two locations using ArcGIS Map 10.7.1 software. The temperature and Air Quality Index (AQI) information were obtained from Environmental Protection Agency (EPA) measurements available through the United States EPA website (http://www.epa.gov/ttn/airs/aqsdatamart). Hourly outdoor temperature and daily AQI data collected by EPA over a span of 5 years (2015-2019) were used in this study. For county-level $PM_{2.5}$ estimates, temporally averaged $PM_{2.5}$ data previously published by Wu et al. (Wu et al., 2020) were used in this study. Briefly, the monthly averages of $PM_{2.5}$ estimates over the entire US were made at 0.1^{o} X 0.1^{o} grid resolution through a combination of satellite-derived estimates, ground-based measurements, and their statistical fusion through a geographically weighted regression model (Donkelaar et al., 2019). This data was further aggregated to the geographical confinement of a county and temporally averaged for the years 2000 – 2016 to obtain a single $PM_{2.5}$ estimate for each county (Wu et al., 2020). We used this average $PM_{2.5}$ data from the past years in the current study. While the exact mechanisms by which $PM_{2.5}$ influences COVID-19 are not fully understood yet, our choice of average $PM_{2.5}$ from past years is motivated by the findings of multiple studies that point to the association of historical exposure of $PM_{2.5}$ to the disease (Gupta et al., 2020; Maleki et al., 2021; Wu et al., 2020).

2.2 Statistical Analyses

2.2.1 K-Means Clustering:

The counties in the NYS were classified into three categories using k-means clustering technique. Partitioning the counties into three disjoint clusters on the basis of infection, death, and fatality was performed to explore any common pattern that might exist among the counties classified within a cluster. For the implementation of the clustering algorithm, the value of k was

set in advance along with the assignment of initial centroid positions for the clusters (Fahim et al., 2006). The algorithm started with the random initialization of the positions of centroids and was followed by two steps. The first step assigned each sample to its nearest centroid. The second step created a new centroid by taking the mean value of all the samples assigned to each previous centroid. The differences between the old and the new centroids were computed, and the algorithm repeated these last two steps until this difference was less than a threshold. The model used Euclidean distance for the calculation of the distance and the threshold considered was 0.0001. In the end, the centroids were fixed, did not move anymore, signifying the convergence criterion for clustering, and resulted in three distinct clusters. Clustering for infection and death was performed using the infection and death rate values from each county. To cluster the counties for fatality, k-means clustering technique was implemented on infection and fatality rate, considering them as two dimensions. Clusters of counties constructed this way were used to study the association with demographic and environmental risk factors.

Table 1. Publicly available data sources used in this study.

Data	Source
Covid-19 cases & deaths	Coronavirus in NY: Cases, maps, charts, and resources (https://www.syracuse.com/coronavirus-ny/)
Population estimates & demographics 2018	US Census Bureau's American Community Survey (ACS) (https://www.census.gov/programs-surveys/acs)
Temperature & air quality index	EPA (http://www.epa.gov/ttn/airs/aqsdatamart)
Nursing homes locations	The Department of Health and Human Services (HHS) (https://www.arcgis.com/home/item.html?id=b3813b2d3a054c378247bf32bcd8d203)
Satellite $PM_{2.5}$ estimates	Air pollution and COVID-19 mortality in the United States, Harvard University (http://github.com/wxwx1993/PM_COVID)

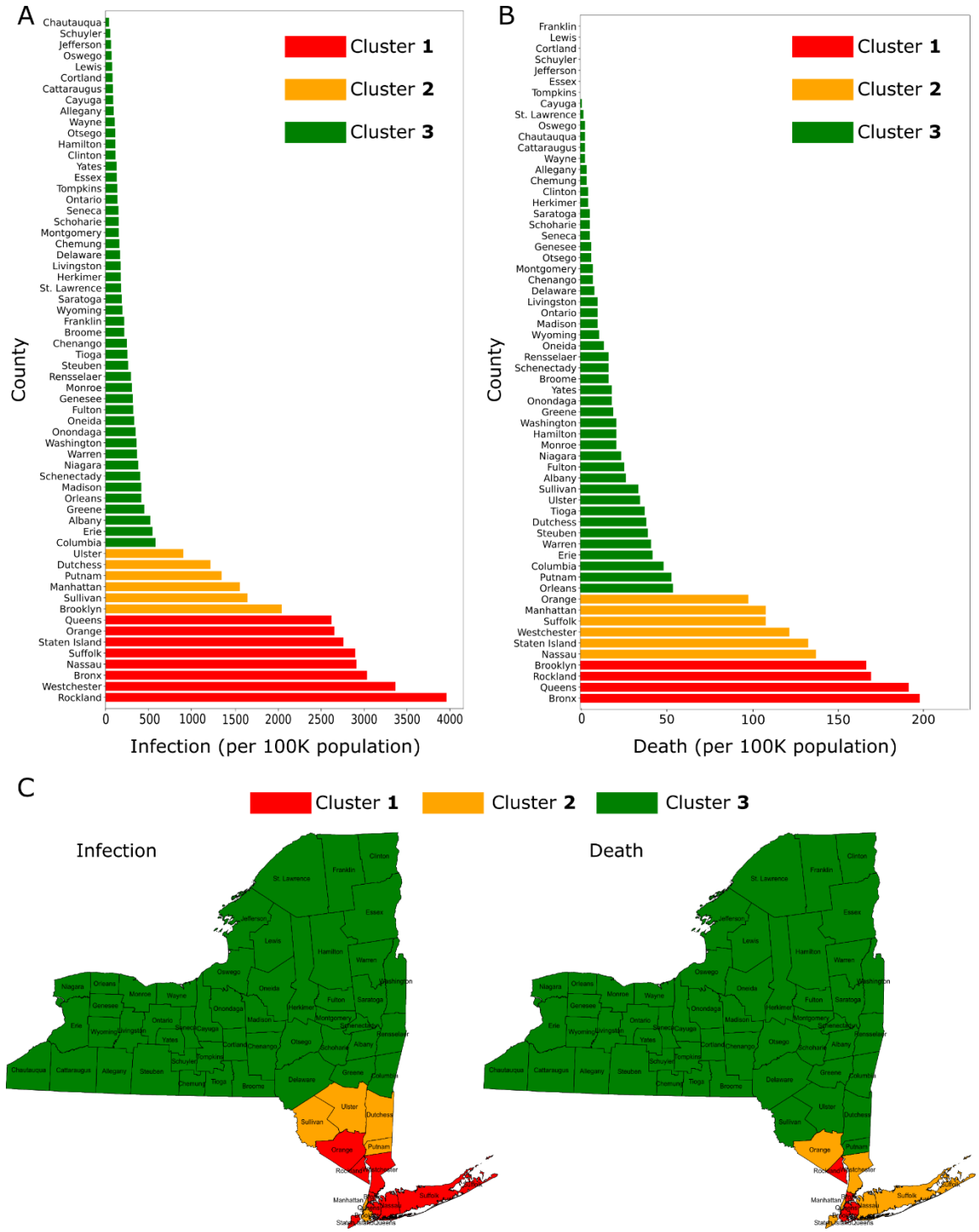


Figure 1. Infection and death from COVID-19 in NYS counties (data till May 16, 2020). (A-B) Infection rates (A) and death rates (B) in individual counties, which are further grouped into three clusters using k-means clustering technique. (Clusters **1, 2,** and **3** represent the counties with high, intermediate, and low infection or death) (C) Maps of NYS showing the locations of counties in each cluster.

184

2.2.2 Tests for Significance:

Statistical comparisons of demographic and environmental variables between the clusters were performed using Kruskal-Wallis (KW) test, a non-parametric equivalent of one-way analysis of variance (ANOVA), since the data were non-normally distributed. Once the KW test statistic was found to be significant, multiple comparisons were conducted using Mann-Whitney U test after making the Bonferroni corrections. All analyses used 2-sided statistical tests and $P < 0.10$ was considered as significant. The Bonferroni correction was set at the significance cut-off value of 0.03.

2.2.3 Autoregressive Integrated Moving Average (ARIMA) Model:

Temperature and AQI time series data from EPA were used to build ARIMA models to obtain predicted estimates of these variables. Data from one representative EPA site for each of the three clusters in each category of infection, death, and fatality were included in the analysis. The models were constructed using time series data from the years 2015-2019. Hourly outdoor temperature data and daily AQI data collected from EPA were first converted to weekly data before using in the model.

In ARIMA model, the future values of a variable are predicted by a linear combination of past values and errors (Hyndman and Athanasopoulos, 2018). The model is often expressed as ARIMA (p, d, q), where p, d, and q represent the order of auto-regression, the degree of trend difference, and the order of moving average, respectively. The model is essentially a combination of three parts: (1) The first part is the auto-regressive model, which uses the linear combination of past values of the variable to forecast the next value and is referred as an AR(p) model, an autoregressive model of order p. (2) The second part is the integrated (I), which is computed by taking the difference between the consecutive observations to make the data stationary. (3) The third part is the moving average (MA) model, referred as MA(q) and equivalent to a regression model that involves past forecast errors as predictors. Augmented Dickey-Fuller (ADF) unit-root test was performed prior to model building to confirm the stationarity of each time series data. Implementing the ARIMA model, predicted time series values with 95% confidence interval were determined for each condition. The model goodness of fit was further evaluated by calculating the Akaike information criterion (AIC). Model predicted values for each cluster within a category were used to compare the temporal pattern of temperature and AQI between the clusters.

2.2.4 Regression Models:

Regression models were built using the data from individual counties of NYS. Three separate models were built where infection, death, and fatality rate were considered as the response variable, while demographic and environmental factors were used as predictor variables for all three models. Variables were first evaluated for normality of distribution by visual inspection of histograms followed by the Shapiro-Wilks test for normality. The univariate method of outlier detection was used to eliminate outliers in the predictors. Correlations between variables were examined by calculating

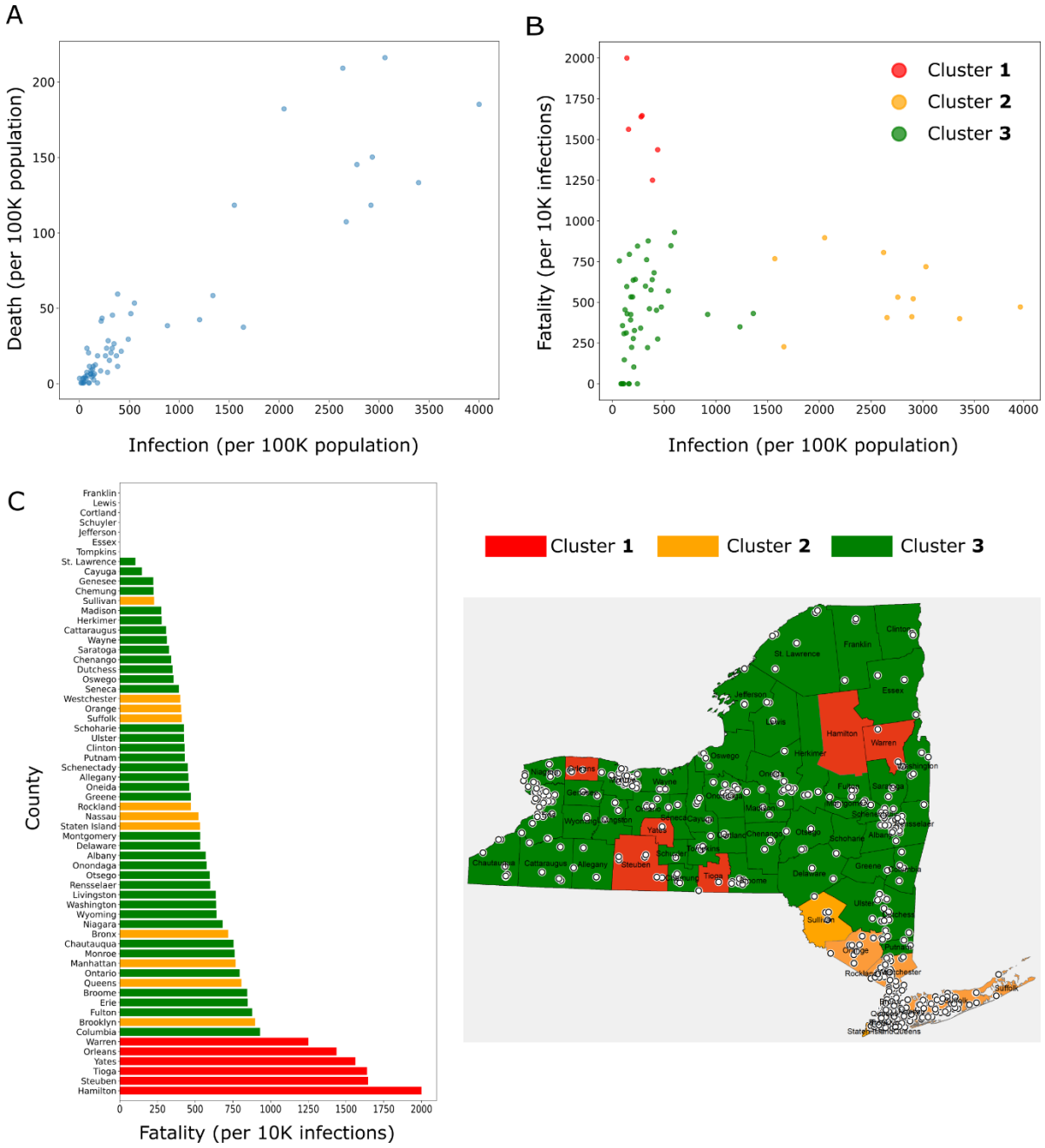


Figure 2. (A) Scatter plot showing the relationship between COVID-19 infection and death rates in NYS counties. (B) Fatality rate is plotted against infection rate for each county; the counties are further grouped for fatality into 3 clusters using two-dimensional k-means clustering method. Cluster **1** includes counties with high fatality and low infection; cluster **2** includes counties with low fatality and high infection; cluster **3** includes counties with low fatality and low infection. (C) Fatality rates of the counties with the clusters indicated by color *(left);* map of NYS showing the locations of counties belonging to each cluster *(right).* Locations of nursing homes are also depicted in the map by white circles.

225

Pearson's correlation coefficients between the predictor and response variables. Multicollinearity between the predictor variables was further examined by computing variance inflation factor (VIF), which measures the inflation in the variances of parameter estimates due to multicollinearity. An upper cut-off value of VIF was set as 5 to minimize the contribution of multicollinearity in our model (Chatterjee and Simonoff, 2013). A stepwise forward selection procedure was implemented to evaluate the contribution of predictor variables in infection, death, and fatality from COVID-19. The forward selection algorithm for stepwise regression starts with an empty model with predictor variables are added sequentially along with the measurement of model accuracy. This process is repeated until all variables are incorporated into the model. The residuals of the regression models were checked for model adequacy and outliers were removed when needed. The goodness of the model is interpreted by the adjusted R^2 value and the contribution of an individual variable is assessed from the order in which the variable was entered in the model. P values of the regression coefficients of the predictor variables were used to assess if their incorporation made a meaningful addition to the model.

All analyses were performed using version 3.6.9 of the Python programming language.

3 Results

3.1 Distribution of COVID-19 in NYS counties

The infections and deaths from COVID-19 in the NYS between March 1 and May 16, 2020, were considered in our analysis. This time window roughly corresponds to the first COVID-19 wave observed in the NYS. To understand the distribution of infections and deaths across the counties within the state, we grouped the counties into three clusters based on each of these variables. Infection and death rates were calculated for all 62 counties in the NYS and then the counties were classified into three clusters using k-means clustering technique. Cluster **1** included counties with a high rate of infection or death, cluster **3** incorporated counties with a low rate, and in cluster **2** the rates were intermediate between the other two clusters (Figure 1A, 1B). For infections, we observed that the cluster **1**, where the infection numbers ranged 2,500-4,000 per 100,000 population, consisted of 8 counties (Rockland, Westchester, Bronx, Nassau, Suffolk, Staten Island, Orange, and Queens), all located in close proximity within the downstate NY (Figure 1C; counties shaded in red). Cluster **2** was formed by 6 counties located near to cluster **1**, namely Ulster, Dutchess, Putnam, Manhattan, Sullivan, and Brooklyn (Figure 1C; counties shaded in yellow). The counties of upstate NY fell in the cluster **3** (Figure 1C; counties shaded in green) where the infection rate was <500 per 100,000 population, well below the other two clusters. The clusters for COVID-19 death showed a similar distribution to infection. Four counties in downstate NY, namely Bronx, Queens, Rockland, and Brooklyn were included in the cluster **1** with death rates ranging from 175 to 200 per 100,000 population (Figure 1B, 1C; counties shaded in red). Of the remaining counties, 6 neighboring counties belonged to cluster **2** (counties shaded in yellow), and the rest of upstate counties were included in the cluster **3** (counties shaded in green) with a death rate <50 per 100,000 population. Thus, clustering the counties followed by visual inspection revealed that higher COVID-19 infections and deaths were from the counties located in downstate NY (Figure 1C).

A similar pattern in the distribution of counties in the clusters for COVID-19 infection and death suggests an association between these two variables, which was confirmed from the scatter plot

(Figure 2A) and a strong positive correlation (Pearson's correlation; r = 0.92, P < 0.0001). The observation suggests that the number of infections in a county is a key determinant for the number of COVID-19 deaths.

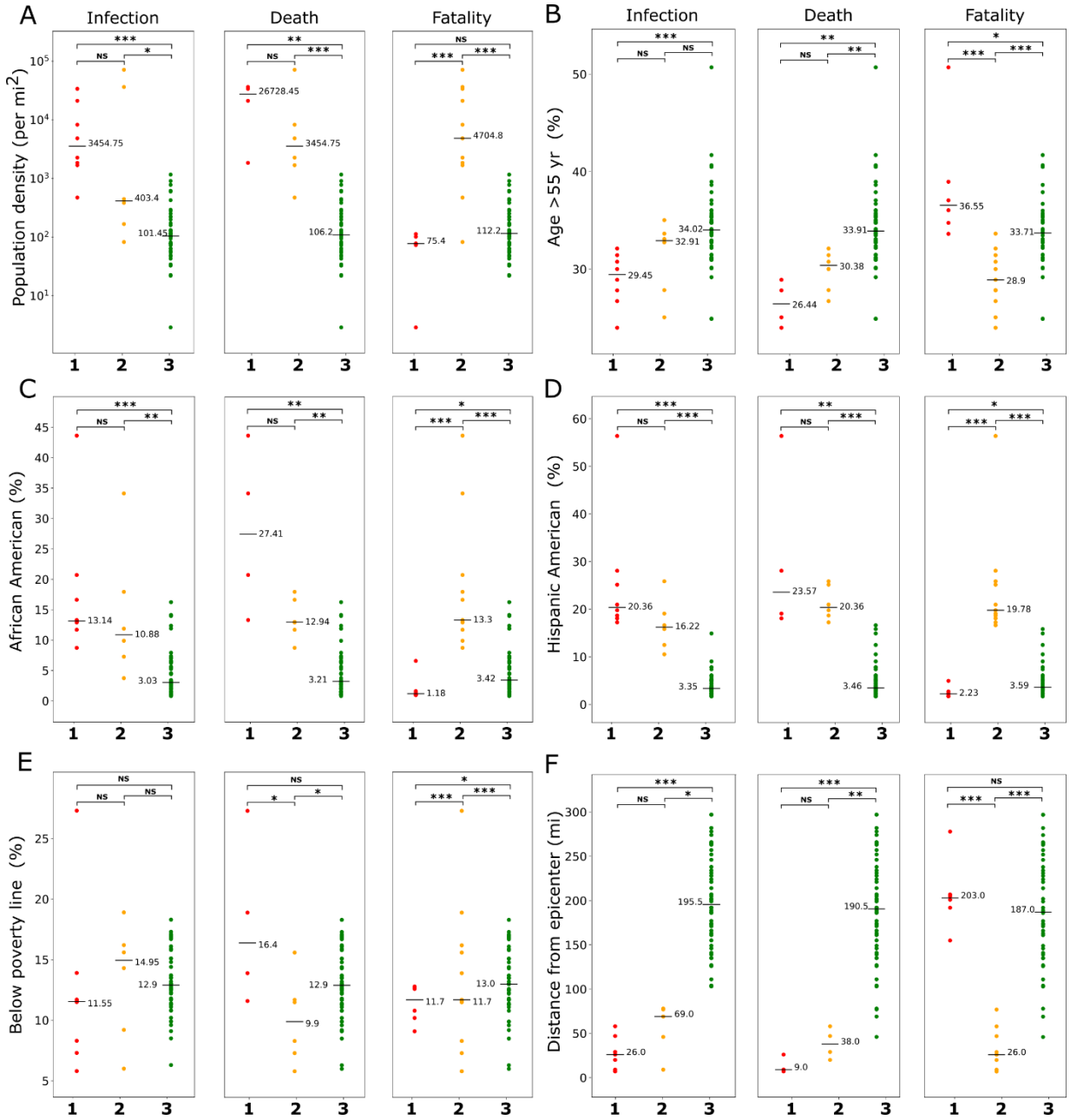


Figure 3. Association of demographic variables with COVID-19. Counties grouped into clusters by infection, death, or fatality were compared for (A) population density, (B) age (> 55 yr), (C) African American population, (D) Hispanic American population, (E) population below poverty line, and (F) distance from epicenter. Horizontal bars represent medians. ***P < 0.01, **P < 0.05, *P < 0.10, NS not significant (Kruskal-Wallis test followed by Mann-Whitney U test with Bonferroni corrections). For infection and death category, clusters **1**, **2**, and **3** represent high, intermediate, and low rates of infection or death; for fatality category,

cluster **1** indicates high fatality and low infection, cluster **2** indicates low fatality and high infection, and cluster **3** indicates low fatality and low infection.

Even though we observed a strong correlation between the infection and death rates, this data does not provide information about the disease fatality, that is the proportion of deaths occurring from infections. When the fatality rate (expressed as deaths per 10,000 infections) was calculated for all counties and plotted against the infection rate, a distinct pattern of relationship between these two variables was found (Figure 2B). We observed that the counties with a high fatality rate had a relatively low infection rate while the counties with high infection rates had a relatively low fatality rate. In accordance, when the counties were divided into three clusters on the basis of infection and fatality rate using a two-dimensional k-means clustering method, we obtained clusters with the following features (Figure 2B): (1) High fatality and low infection rate (cluster **1**); (2) high infection and low fatality rate (cluster **2**); (3) low infection and low fatality rate (cluster **3**). Interestingly, the locations of the counties included in cluster **1** (Hamilton, Steuben, Tioga, Yates, Orleans, and Warren) were distributed across the NYS (Figure 2C; counties shaded in red). The counties in cluster **2** (Figure 2C; counties shaded in yellow) were all in proximity and located near the NYC, although in terms of the fatality rate, they were interspersed with cluster **3** (Figure 2C; counties shaded in green). Since high fatality from COVID-19 is observed among nursing residents (Rada, 2020), we also checked whether the distribution of nursing homes has a contribution to the variation of fatality observed between NYS counties. Mapping the nursing homes in individual counties, we did not find an apparent relationship between counties with high fatality and increased density of nursing homes (Figure 2C). These results suggest that various risk factors for COVID-19 have a differential contribution on infection and fatality.

3.2 Impact of demographic factors on COVID-19

Multiple studies have shown the association of demographic variables with COVID-19 infection and outcome (Goldstein and Lee, 2020; Karmakar et al., 2021; Perone, 2021; Sorci et al., 2020). To study how they vary across the clusters of NYS counties that we constructed on the basis of COVID-19 infection, death, and fatality, we selected five well-known demographic risk factors namely, population density, age (percentage of people with age above 55 yr), ethnicity (percentage of African American and Hispanic American population), and poverty (percentage of the population with income below poverty line). Additionally, we considered the distance from the disease epicenter, measured as the distance of a county from Manhattan in NYC. Figure 3 shows these variables plotted against counties organized in three clusters as described in the previous section. Each variable demonstrated a characteristic pattern of distribution within the clusters. KW test followed by multiple comparisons was further performed to calculate the statistical difference.

The trends for most demographic variables followed a similar pattern for infection and death clusters except for poverty. For population density and ethnicity (African American or Hispanic American) median values showed a decreasing trend from cluster **1** to cluster **3**, while an opposite trend was observed for age and distance from the epicenter (Figure 3). Furthermore, the difference between clusters **1** and **2** was not significant for these variables but their difference

with cluster **3** was found to be significant (except between clusters **2** and **3** for age). Interestingly, for the percentage of the population below the poverty line, the highest median value was observed in cluster **2** for the infection group, and in cluster **1** for the death group; the median values in cluster **3** was intermediate for both infection and death groups, thus, suggesting a role of poverty on COVID-19 infection and death cannot be explained through simple association.

The clusters in the fatality group demonstrated a highly distinct pattern of association with demographic variables (Figure 3). For all variables except poverty, the median values of cluster **3** (low fatality and low infection) were found to be intermediate between the median values of cluster **1** (high fatality and low infection) and cluster **2** (low fatality and high infection). Specifically, for the percentage of population with age over 55 yr, the highest median value was observed in cluster **1** (36.5%) in comparison to 28.9% and 33.7% observed in cluster **2** and cluster **3**, respectively. For population density and ethnicity, the trend was opposite with cluster **2** and cluster **1** showing the highest and lowest median values among the clusters, respectively (Figure 3). These findings indicate that high fatality and infections are associated with different sets of demographic risk factors. Overall, the analysis suggests a potential role of demographic structure towards the extent of observed infection, death, and fatality from COVID-19.

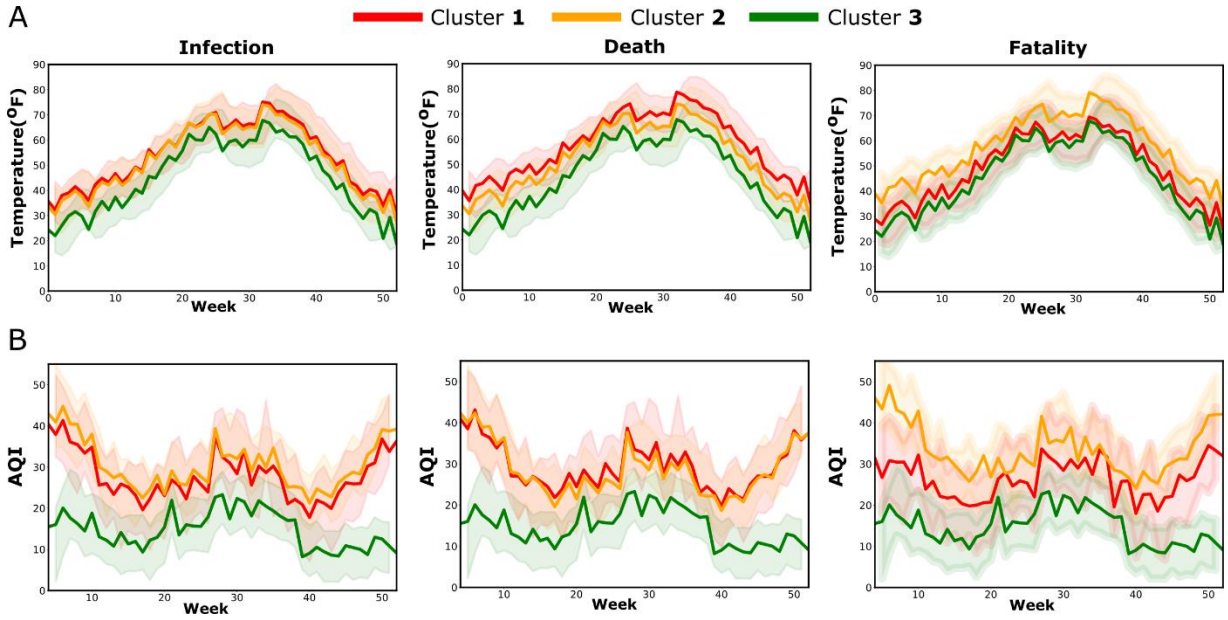


Figure 4. ARIMA time series analysis of temperature (A) and air AQI (B) from weekly EPA data (2015-2019). Predicted values with 95% confidence band for one year, starting from January are shown in the plots. EPA sites for each cluster were located in a representative county belonging to that cluster. For infection and death categories, clusters **1**, **2**, and **3** represent high, intermediate, and low rates of infection or death; for fatality category, cluster **1** indicates high fatality and low infection, cluster **2** indicates low fatality and high infection, and cluster **3** indicates low fatality and low infection. Abbreviations: ARIMA, autoregressive integrated moving average; AQI, air quality index.

3.3 Impact of environmental factors on COVID-19

Since several recent studies have shown an association of environmental factors such as air pollution and temperature on COVID-19 transmission and severity (Li et al., 2020; Wu et al., 2020), we investigated whether such association could be observed across the clusters of NYS counties. Furthermore, we hypothesized chronic exposure to have a stronger impact than acute exposure. This prompted us to select one EPA site representative for each cluster and collect temperature and AQI data for the years 2015-2019. To compare the variables between the sites and find out any differences throughout the year or any specific period of the year, ARIMA models were constructed from weekly time series data. Figure 4 shows ARIMA models of predicted values with 95% confidence bands for temperature and AQI. The AIC values of the models for all conditions were low and comparable (range 289.68 – 303.11), confirming the model robustness. The model predicted temperatures showed a similar pattern for all three clusters in all three categories of infection, death, and fatality with values reaching a peak during the summer months of June-August. Although the predicted temperature for cluster **3** in the infection and the death categories were slightly lower than the other two clusters, there was a considerable overlap of the confidence bands, thus an association the clusters with temperature could not be inferred (Figure 4A). Similarly, the confidence bands of the models for temperature in the fatality clusters also demonstrated a substantial overlap. In contrast to temperature, the model predicted AQI values demonstrated a larger separation between the clusters (Figure 4B). In the infection and death groups, AQI values for cluster **3** were substantially lower than the values for clusters **1** and **2**, and the differences were more prominent during the winter months with separation of the confidence bands. In the fatality group, the highest AQI values were observed in the cluster **2**, which corresponds to high infection but low fatality, and the lowest AQI values were observed in the cluster **3**, corresponding to low infection and low fatality. Thus, the analysis of EPA data suggests COVID-19 in NYS is linked to poor air quality but not with outdoor temperature.

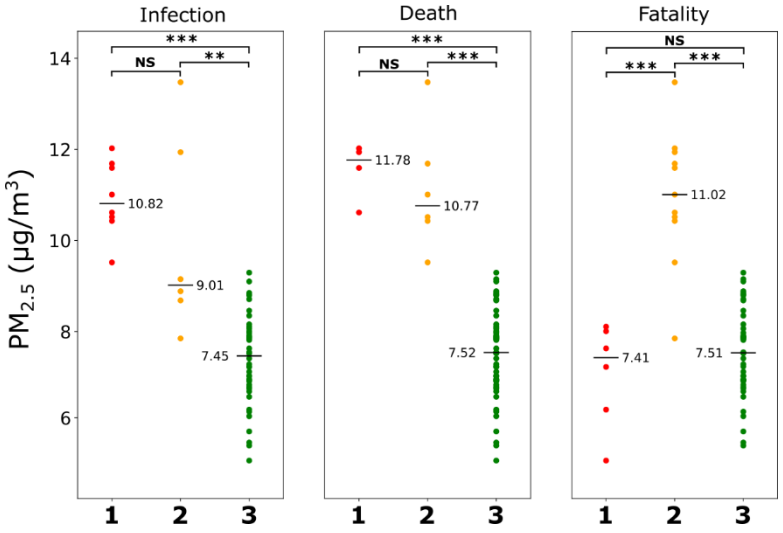


Figure 5. Temporally averaged $PM_{2.5}$ estimates from NYS counties for the years 2000-2016 are compared between clusters based on COVID-19 infection, death, and fatality. Horizontal lines represent median values. ***P < 0.01, **P < 0.05, *P < 0.10, NS not

significant (Mann-Whitney U test with Bonferroni corrections). For infection and death categories, clusters **1**, **2**, and **3** represent high, intermediate, and low rates of infection or death; for fatality category, cluster **1** indicates high fatality and low infection, cluster **2** indicates low fatality and high infection, and cluster **3** indicates low fatality and low infection.

Although EPA measurements provide an accurate estimate of the air quality, data at the resolution of individual counties are not available due to the relatively few EPA sampling sites across the NYS. Therefore, to capture the variation of air quality across the counties, we used temporally averaged $PM_{2.5}$ estimates from satellite data and ground-based measurements over a time period of 2000-2016 (Wu et al., 2020). When the $PM_{2.5}$ values from the counties were compared between the COVID-19 clusters for infection, death, and fatality, the pattern corroborated well with the observations from EPA data (Figure 5). For COVID-19 infection and death, $PM_{2.5}$ values of counties in cluster **3** were significantly lower than the counties in clusters **1** and **2**, with no significant difference observed between the latter two. Similar to the findings with EPA data, in the fatality group, the $PM_{2.5}$ of counties in cluster **2** was significantly higher than in cluster **1** and **3**. These findings demonstrate the association of $PM_{2.5}$ with COVID-19 infection and death in the NYS.

3.4 Contributions of risk factors on COVID-19 infection, death, and fatality

Since clustering of counties based on COVID-19 infection, death, or fatality demonstrated a distinct pattern of association with demographic or environmental risk factors, we wanted to further elucidate the contribution of these variables on the specific aspects of COVID-19 burden. Six risk factors namely, age above 55 yr, ethnicity (African American and Hispanic American population), poverty, population density, distance from the epicenter, and $PM_{2.5}$ were considered as predictor variables, and multivariate regression model with forward "stepwise" selection was used for analysis. Three separate models were constructed using infection, death, and fatality rate as dependent variables to understand the relative contribution of the risk factors for each of these outcomes. Rockland county was excluded from the models as it was identified as an outlier while performing residual analyses of the regression output.

Multicollinearity among the predictor variables can lead to unstable and unreliable estimates of regression coefficients, reducing the power of the regression model. Therefore, before their incorporation in the regression models, we checked for multicollinearity. The correlation matrices in Figure 6A show the Pearson's correlations coefficients among variable pairs. A strong positive correlation was found between ethnicity and $PM_{2.5}$ ($r = 0.81$, $P < 0.001$), while moderate positive correlations were observed between population density and $PM_{2.5}$, ($r = 0.69$, $P < 0.001$) or ethnicity ($r = 0.67$, $P < 0.001$). Additionally, ethnicity and $PM_{2.5}$ demonstrated a strong positive correlation with infection and death, while the distance from the epicenter held a strong negative correlation with these dependent variables. Interestingly, such strong correlations were not observed for fatality, the third dependent variable.

The existence of multicollinearity among predictor variables prompted us to calculate the VIF for each variable to assess their suitability for inclusion in the regression model. We found that VIFs of all variables were lower than the acceptable cut-off value of 5, except for ethnicity when

infection or death was used as dependent variables. This implies that ethnicity is not an independent predictor for infection and death, and therefore, was excluded in the regression models for these two variables. The models revealed distinct contributions of the predictor variables to infection, death, and fatality rates (Figure 6B). $PM_{2.5}$ and distance from the epicenter were found to be the two most important predictors for infection and death. For infection rate, distance from the epicenter was the strongest predictor with a highly significant regression coefficient ($P<0.001$, Figure 6C) and generated an adjusted R^2 value of 0.60 when considered as a sole contributor of the model (Figure 6B); the adjusted R^2 value increased to 0.71 following the inclusion of $PM_{2.5}$ in the model, which also had a significant regression coefficient ($P<0.001$; Figure 6B, C). Regression coefficients of population density, age, and poverty emerged as not significant ($P>0.05$) and their addition to the regression model only marginally increased the adjusted R^2 value 0.74. Similar to infection, distance from the epicenter and $PM_{2.5}$ were two major predictors for the death rate, however, $PM_{2.5}$ was the strongest among them contributing to an adjusted R^2 value of 0.69. The value increased to 0.73 following the inclusion of distance from the epicenter in the model but did not change further upon the addition of other variables.

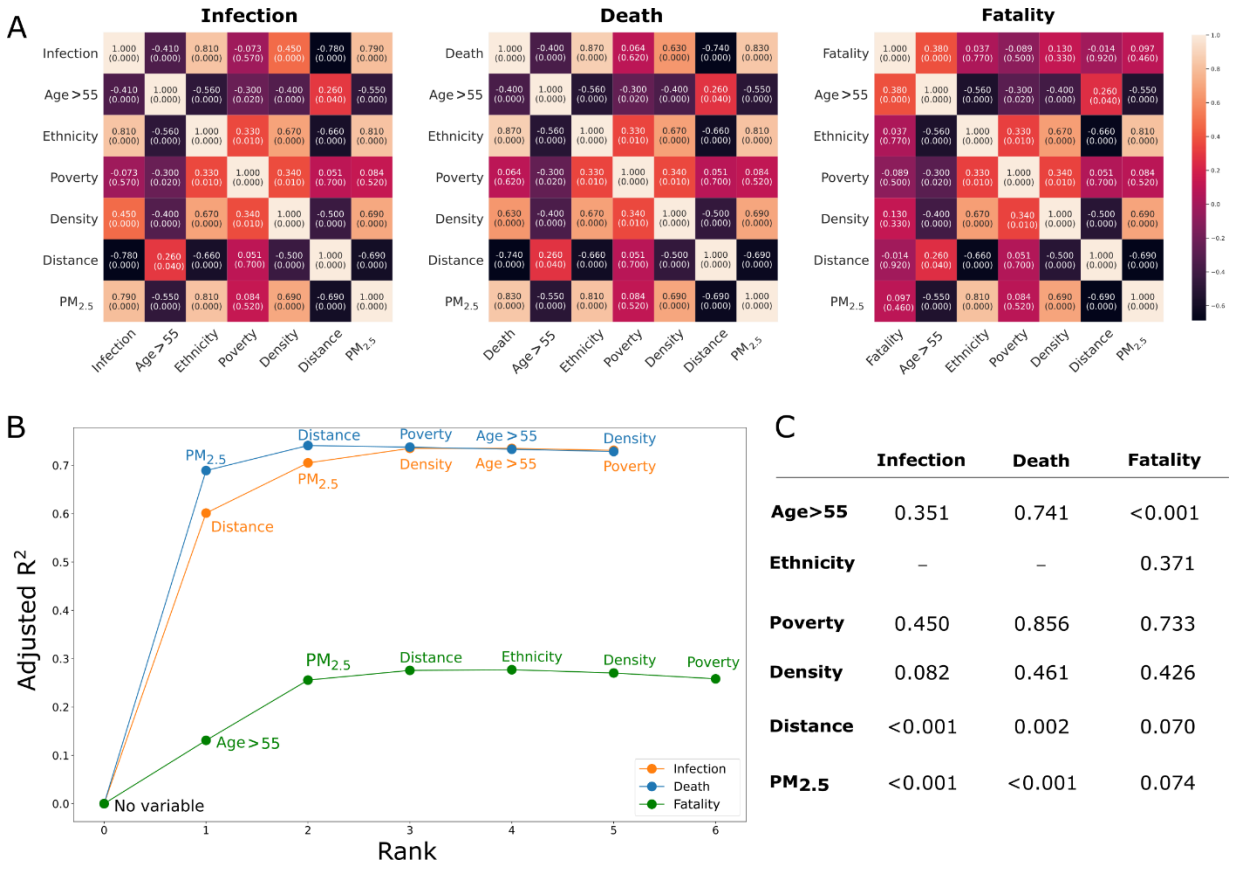


Figure 6. Regression analysis to assess relative contribution of risk factors on COVID-19 infection, death, and fatality. (A) Correlation matrices between demographic risk variables and $PM_{2.5}$ for infection, death, and fatality. Pearson's correlations coefficients between the variables are shown with corresponding P values are mentioned in the parentheses. (B) Stepwise regression model with forward selection for infection, death, and fatality. (Note that

ethnicity is excluded from the models for infection and death due to the existence of strong multicollinearity). (C) Table showing the P values of the regression coefficients from the analysis. Abbreviations used: Density, population density (per mi^2); Ethnicity, African American and Hispanic American (%); Poverty, population below poverty line (%); Distance, distance from the epicenter (mi).

Unlike the models for infection and death rates, age over 55 yr was found to be the strongest predictor in the model for fatality rate, (Figure 6B, C). $PM_{2.5}$ was the second major predictor in the model, although the relatively high P-value (0.074) of the regression coefficient suggests a weaker link with fatality. It should also be noted that the adjusted R^2 value of the final model for fatality was 0.26, indicating that the goodness of model fit is much lower than the other two models for infection and death. Together, the regression analysis helped to delineate the risk factors with major contributions on specific aspects of the COVID-19 burden.

4 Discussion

Our analysis has shown a wide heterogeneity in the infection, death, and fatality rates from COVID-19 among the counties in the NYS during the first pandemic wave. Infection was found to be strongly correlated with death but not with fatality. By grouping the counties into clusters, we show the association of multiple demographic factors and air quality with infection, death, and fatality from COVID-19. Specifically, $PM_{2.5}$, population density, and proportion of African American or Hispanic American population demonstrated a positive association with infection and death, while the distance from the disease epicenter showed a negative association. In contrast, higher fatality from the disease was primarily associated with a higher proportion of the population aged above 55 yr. Furthermore, regression analysis has identified the major contributors among these risk factors for infection, death, and fatality. These results could help to better understand the impact of environmental and demographic factors on COVID-19 in NYS.

Our analysis shows an association of $PM_{2.5}$ with infection and death, a potential but weaker link to fatality. This finding is in agreement with studies focused on the role of outdoor air pollution on COVID-19 transmission and health outcomes (Benmarhnia, 2020; Gupta et al., 2020; Lolli et al., 2020; Pozzer et al., 2020; Wu et al., 2020). $PM_{2.5}$ has been reported to be positively correlated with both increased COVID-19 transmission and fatality. Comorbid conditions such as respiratory and cardiovascular illnesses that are associated with chronic exposure to higher $PM_{2.5}$ are thought to aggravate the illness from virus infection, posing a higher risk of death (Benmarhnia, 2020; Pozzer et al., 2020; Wu et al., 2020). Additional mechanisms proposed for increased SARS-CoV-2 transmission by $PM_{2.5}$ include the particulate matters serving as a transport vector and their ability to increase the susceptibility to infection by inducing lung inflammation (Maleki et al., 2021). It is to be noted that although the average $PM_{2.5}$ values in the NYS meet the safe limit (< 12 $\mu g/m^3$) set by EPA (Jin et al., 2019), an association of $PM_{2.5}$ level with adverse COVID-19 outcome can be clearly discerned. We attribute this apparent discrepancy to the existence of potential hot spots where the exposure $PM_{2.5}$ could be much higher than the average. Although our analysis considers chronic exposure to $PM_{2.5}$ (average values from 2000-2016) to better capture the long-term health effects(Wu et al., 2020), a recent

study showed no significant difference in air quality in the NYC area after lockdown (Zangari et al., 2020), and thus would also reasonably reflect the exposure to $PM_{2.5}$ during the period of investigation. Also, our work focuses on $PM_{2.5}$ alone, however, other air pollutants, especially NO_2 and O_3 are reported to be associated with higher COVID-19 spread and fatality (Adhikari and Yin, 2020; Copat et al., 2020; Liang et al., 2020). Commonly generated by anthropogenic activities such as traffic, NO_2 is a well-known inducer of lung inflammation and can synergistically act with $PM_{2.5}$ to increase the adverse impact of COVID-19 (Hesterberg et al., 2009; Huang et al., 2012). NO_2 is also a source for O_3 formation, which is further facilitated by higher temperature and $PM_{2.5}$ (Zhang et al., 2019). O_3 is a strong oxidant and exposure to O_3 is known to cause or aggravate respiratory and cardiovascular diseases, and older adults are shown to be more vulnerable to the adverse effects (Day et al., 2018; Zhang et al., 2019). Thus, apart from independent effects, the interactions of NO_2 and O_3 with $PM_{2.5}$ could be important in the context of COVID-19 and needs to be investigated in the future study. Outdoor air temperature, the other environmental parameter we examined in this study did not show any association with COVID-19 cases; both positive and negative association of temperature with COVID-19 infection have been reported in the literature, and a recent systematic review concluded data-related and methodological issues including inherent uncertainties of the data, inappropriate controlling for confounding parameters, and short periods of investigation underlie such conflicting results (Dong et al., 2021)(Lolli et al., 2020). Thus, to better understand the influence of temperature on COVID-19, future studies should be conducted with time-resolved data for a longer period and taking appropriate measures for confounding variables.

We observed a strong correlation between $PM_{2.5}$ and the percentage of the population belonging to African American or Hispanic American ethnicity ($r = 0.81$, $P < 0.001$), suggesting that people from these ethnicities are exposed to a higher level of $PM_{2.5}$ than the average population. Multiple studies have concluded that these two ethnicities in the US are at disproportionately higher risk of COVID-19 (Cordes and Castro, 2020; Li et al., 2020; Martinez et al., 2020; Yancy, 2020). Factors related to socioeconomic inequities such as the greater risk of virus exposure from professional demand or living in crowded accommodation, higher prevalence of chronic comorbidity, and restricted access to healthcare are thought to underlie such differences (Patel et al., 2020). Our results suggest that exposure to air pollution could be a contributor to further increase this disparity. Low socioeconomic status is thought to pose a greater risk for COVID-19 exposure (Yancy, 2020); however, for NYS, we observed counties in cluster **3** for infection and death rates to have a higher proportion of people living below the poverty line than the counties from other two clusters. That cluster **3** counties have a relatively lower risk from other demographic and environmental variables could explain this apparent discrepancy. Indeed, when considering the population of NYC alone, poverty and COVID-19 are found to be positively correlated (Cordes and Castro, 2020).

Two demographic variables, distance from the epicenter and age above 55 years, came out as major contributors in our regression models. However, their influences on COVID-19 were distinct. The distance of counties from the disease epicenter was inversely related to infection and death, alone accounting for 60% of variation in the regression model of infection. This finding is not unusual as the disease spread would be facilitated by the population mobility with the highest effect on the neighboring regions. We also observed a strong correlation between

infection and death rates ($r = 0.92$, $P < 0.0001$), corroborating their association with a similar set of risk factors. In contrast, a poor correlation was observed between infection and fatality, and the regression model for fatality revealed age over 55 years to be the most significant independent variable. Fatality from COVID-19 depends on the health of patients where age plays a crucial role (Mesas et al., 2020; Richardson et al., 2020). Increased risk of the aged population to complications and death from COVID-19 is observed across the world, and multiple factors including the existence of chronic comorbid conditions and a weaker immune system are thought to underlie such vulnerability (Mesas et al., 2020). The association of distinct sets of risk factors for death and fatality suggests that they should be considered as separate metrics for the COVID-19 burden for the development of preventive or mitigative strategies.

Grouping the counties into clusters not only helped to visualize how NYS counties are impacted by specific COVID-19 adversities but also allowed easier comparison of their association with various demographic and environmental risk variables. While the regression models have further helped to identify the risk variables that have major contributions to specific aspects of disease impact, it should also be noted that the adjusted R^2 value of the regression model for fatality is substantially lower than the models for infection and death (0.26 *vs.* 0.74 and 0.73). This difference suggests the possibility of missing key variables in the model for COVID-19 fatality that needs to be identified and incorporated in the future model. Such variables could potentially be measures pertaining to the outcome of an infected individual, including the availability and access to healthcare resources, vaccination, and awareness for early diagnosis and treatment.

Conclusions

In this work, we analyzed the association of multiple demographic and environmental factors with the COVID-19 burden in NYS during the first pandemic wave. Clustering the counties based on COVID-19 infection or death revealed their segregation by geographical location with clusters located farther away from NYC showing lower infection or death. In contrast, counties grouped in the cluster for high disease fatality were distributed across the NYS and were different from those having high infection and death rates. The clustered counties showed a prominent association with demographic variables and $PM_{2.5}$ but the patterns of association for infection and death were distinct than for fatality. Clusters with high infection and death were found to have higher $PM_{2.5}$, higher population density, a higher proportion of African Americans and Hispanic Americans, and were closer to the disease epicenter, while the cluster with higher fatality had a higher proportion of population aged above 55 yr. Stepwise regression models built on county data further showed that $PM_{2.5}$ and the distance from the epicenter are two major contributors for infection and death, while advanced age makes the strongest contribution to fatality. Although our study is confined to counties within the NYS, we observed prominent differences in the distribution of infection and fatality along with an association with distinct sets of demographic and environmental risk variables. The US being a country with a vast size, have considerable heterogeneity between states in terms of social and cultural practices, public health policies, access to healthcare, and general awareness of COVID-19, all of which could have a significant impact on the absolute magnitude of COVID-19 burden; however, we expect that the variables considered in this work would still have similar effects as observed for the NYS, and

thus, our results can provide key insight on the contribution of demographic and environmental factors on the disease landscape in these states. Additionally, a similar modeling approach could be utilized in future studies to include additional relevant variables in the analysis to understand their contribution to the disease. With strong anthropogenic contributions to the environment in modern societies, our findings suggest the need for critical consideration of both demographic and environmental variables when predicting the impact of COVID-19 or developing preventive or mitigative strategies to control the disease.

Funding information

Not applicable.

Declaration of conflicts of interest

The authors declared that they have no conflicts of interest.

Acknowledgments

Vijay Kumar acknowledges the support from US-Pakistan Knowledge Corridor PhD Scholarship Program under Higher Education Commission, Pakistan. Bridget Wangler thanks the Clarkson University Honors Program for their support.

References

Adhikari A, Yin J. Short-Term Effects of Ambient Ozone, PM2.5, and Meteorological Factors on COVID-19 Confirmed Cases and Deaths in Queens, New York. Int J Environ Res Public Health 2020; 17: 4047.

Arif M, Sengupta S. Nexus between population density and novel coronavirus (COVID-19) pandemic in the south Indian states: A geo-statistical approach. Environment, Development and Sustainability 2020: 1-29.

Auger KA, Shah SS, Richardson T, Hartley D, Hall M, Warniment A, et al. Association Between Statewide School Closure and COVID-19 Incidence and Mortality in the US. JAMA 2020; 324: 859-870.

Baldwin R, Di Mauro BW. Economics in the time of COVID-19: A new eBook: CEPR Press, 2020.

Bashir MF, Ma BJ, Bilal, Komal B, Bashir MA, Tan DJ, et al. Correlation between climate indicators and COVID-19 pandemic in New York, USA. Science of the Total Environment 2020; 728: 138835.

Benmarhnia T. Linkages Between Air Pollution and the Health Burden From COVID-19: Methodological Challenges and Opportunities. American Journal of Epidemiology 2020; 189: kwaa148.

Brook RD, Rajagopalan S, Pope CA, Brook JR, Bhatnagar A, Diez-Roux AV, et al. Particulate Matter Air Pollution and Cardiovascular Disease. Circulation 2010; 121: 2331-2378.

Chatterjee S, Simonoff JS. Handbook of regression analysis. Vol 5: John Wiley & Sons, 2013.

Chauhan AJ, Johnston SL. Air pollution and infection in respiratory illness. Br Med Bull 2003; 68: 95-112.

Chen JT, Krieger N. Revealing the Unequal Burden of COVID-19 by Income, Race/Ethnicity, and Household Crowding: US County Versus Zip Code Analyses. J Public Health Manag Pract 2021; 27 Suppl 1, COVID-19 and Public Health: Looking Back, Moving Forward: S43-S56.

Copat C, Cristaldi A, Fiore M, Grasso A, Zuccarello P, Signorelli SS, et al. The role of air pollution (PM and NO2) in COVID-19 spread and lethality: A systematic review. Environ Res 2020; 191: 110129.

Copiello S, Grillenzoni C. The spread of 2019-nCoV in China was primarily driven by population density. Comment on "Association between short-term exposure to air pollution and COVID-19 infection: Evidence from China" by Zhu et al. Science of The Total Environment 2020; 744: 141028.

Cordes J, Castro MC. Spatial analysis of COVID-19 clusters and contextual factors in New York City. Spatial and Spatio-temporal Epidemiology 2020; 34: 100355.

Day DB, Clyde MA, Xiang J, Li F, Cui X, Mo J, et al. Age modification of ozone associations with cardiovascular disease risk in adults: a potential role for soluble P-selectin and blood pressure. J Thorac Dis 2018; 10: 4643-4652.

Dong ZM, Fan XR, Wang J, Mao YX, Luo YY, Tang S. Data-related and methodological obstacles to determining associations between temperature and COVID-19 transmission. Environmental Research Letters 2021; 16: 034016.

Donkelaar Av, Martin RV, Li C, Burnett RT. Regional Estimates of Chemical Composition of Fine Particulate Matter Using a Combined Geoscience-Statistical Method with Information from Satellites, Models, and Monitors. Environmental Science & Technology 2019; 53: 2595-2611.

Fahim AM, Salem AM, Torkey FA, Ramadan MA. An efficient enhanced k-means clustering algorithm. Journal of Zhejiang University-SCIENCE A 2006; 7: 1626-1633.

Feng C, Li J, Sun W, Zhang Y, Wang Q. Impact of ambient fine particulate matter (PM2.5) exposure on the risk of influenza-like-illness: a time-series analysis in Beijing, China. Environ Health 2016a; 15: 17.

Feng S, Gao D, Liao F, Zhou F, Wang X. The health effects of ambient PM2.5 and potential mechanisms. Ecotoxicol Environ Saf 2016b; 128: 67-74.

Goldstein JR, Lee RD. Demographic perspectives on the mortality of COVID-19 and other epidemics. Proc Natl Acad Sci U S A 2020; 117: 22035-22041.

Guan WJ, Zheng XY, Chung KF, Zhong NS. Impact of air pollution on the burden of chronic respiratory diseases in China: time for urgent action. Lancet 2016; 388: 1939-1951.

606 Gupta A, Bherwani H, Gautam S, Anjum S, Musugu K, Kumar N, et al. Air pollution
607 aggravating COVID-19 lethality? Exploration in Asian cities using statistical models.
608 Environment, Development and Sustainability 2020: 1-10.
609 Hesterberg TW, Bunn WB, McClellan RO, Hamade AK, Long CM, Valberg PA. Critical review
610 of the human data on short-term nitrogen dioxide (NO2) exposures: evidence for NO2
611 no-effect levels. Crit Rev Toxicol 2009; 39: 743-81.
612 Hopke PK, Croft D, Zhang W, Lin S, Masiol M, Squizzato S, et al. Changes in the acute
613 response of respiratory diseases to PM2.5 in New York State from 2005 to 2016. Sci
614 Total Environ 2019; 677: 328-339.
615 Huang YC, Rappold AG, Graff DW, Ghio AJ, Devlin RB. Synergistic effects of exposure to
616 concentrated ambient fine pollution particles and nitrogen dioxide in humans. Inhal
617 Toxicol 2012; 24: 790-7.
618 Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice: OTexts, 2018.
619 Jin XM, Fiore AM, Civerolo K, Bi JZ, Liu Y, van Donkelaar A, et al. Comparison of multiple
620 PM2.5 exposure products for estimating health benefits of emission controls over New
621 York State, USA. Environmental Research Letters 2019; 14: 084023.
622 Karmakar M, Lantz PM, Tipirneni R. Association of Social and Demographic Factors With
623 COVID-19 Incidence and Death Rates in the US. JAMA Netw Open 2021; 4: e2036462.
624 Lee VJ, Chiew CJ, Khong WX. Interrupting transmission of COVID-19: lessons from
625 containment efforts in Singapore. J Travel Med 2020; 27.
626 Li AY, Hannah TC, Durbin JR, Dreher N, McAuley FM, Marayati NF, et al. Multivariate
627 Analysis of Black Race and Environmental Temperature on COVID-19 in the US. Am J
628 Med Sci 2020; 360: 348-356.
629 Liang D, Shi L, Zhao J, Liu P, Sarnat JA, Gao S, et al. Urban Air Pollution May Enhance
630 COVID-19 Case-Fatality and Mortality Rates in the United States. The Innovation 2020;
631 1: 100047.
632 Liu J, Zhou J, Yao J, Zhang X, Li L, Xu X, et al. Impact of meteorological factors on the
633 COVID-19 transmission: A multi-city study in China. Science of The Total Environment
634 2020; 726: 138513.
635 Lolli S, Chen YC, Wang SH, Vivone G. Impact of meteorological conditions and air pollution
636 on COVID-19 pandemic transmission in Italy. Sci Rep 2020; 10: 16213.
637 Lusignan Sd, Dorward J, Correa A, Jones N, Akinyemi O, Amirthalingam G, et al. Risk factors
638 for SARS-CoV-2 among patients in the Oxford Royal College of General Practitioners
639 Research and Surveillance Centre primary care network: a cross-sectional study. The
640 Lancet Infectious Diseases 2020; 20: 1034-1042.
641 Maleki M, Anvari E, Hopke PK, Noorimotlagh Z, Mirzaee SA. An updated systematic review on
642 the association between atmospheric particulate matter pollution and prevalence of
643 SARS-CoV-2. Environmental Research 2021; 195: 110898.

644 Martinez DA, Hinson JS, Klein EY, Irvin NA, Saheed M, Page KR, et al. SARS-CoV-2
645 Positivity Rate for Latinos in the Baltimore-Washington, DC Region. JAMA 2020; 324:
646 392-395.
647 Mesas AE, Cavero-Redondo I, Alvarez-Bueno C, Sarria Cabrera MA, Maffei de Andrade S,
648 Sequi-Dominguez I, et al. Predictors of in-hospital COVID-19 mortality: A
649 comprehensive systematic review and meta-analysis exploring differences by age, sex
650 and health conditions. PLoS One 2020; 15: e0241742.
651 Miller LE, Bhattacharyya R, Miller AL. Data regarding country-specific variability in Covid-19
652 prevalence, incidence, and case fatality rate. Data in Brief 2020; 32: 106276.
653 Monmonier M, Giordano A. GIS in New York State county emergency management offices:
654 User assessment. Applied Geographic Studies 1998; 2: 95-109.
655 Patel JA, Nielsen FBH, Badiani AA, Assi S, Unadkat VA, Patel B, et al. Poverty, inequality and
656 COVID-19: the forgotten vulnerable. Public Health 2020; 183: 110-111.
657 Perone G. The determinants of COVID-19 case fatality rate (CFR) in the Italian regions and
658 provinces: An analysis of environmental, demographic, and healthcare factors. Science of
659 The Total Environment 2021; 755: 142523.
660 Pozzer A, Dominici F, Haines A, Witt C, Munzel T, Lelieveld J. Regional and global
661 contributions of air pollution to risk of death from COVID-19. Cardiovasc Res 2020;
662 116: 2247-2253.
663 Pradhan A, Olsson PE. Sex differences in severity and mortality from COVID-19: are males
664 more vulnerable? Biol Sex Differ 2020; 11: 53.
665 Rada AG. Covid-19: the precarious position of Spain's nursing homes. BMJ 2020; 369: m1554.
666 Reichberg SB, Mitra PP, Haghamad A, Ramrattan G, Crawford JM, Northwell C-RC, et al.
667 Rapid Emergence of SARS-CoV-2 in the Greater New York Metropolitan Area:
668 Geolocation, Demographics, Positivity Rates, and Hospitalization for 46 793 Persons
669 Tested by Northwell Health. Clin Infect Dis 2020; 71: 3204-3213.
670 Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, et al.
671 Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients
672 Hospitalized With COVID-19 in the New York City Area. JAMA 2020; 323: 2052-2059.
673 Rocklöv J, Sjödin H. High population densities catalyze the spread of COVID-19. Journal of
674 Travel Medicine 2020; 27.
675 Sarkodie SA, Owusu PA. Global assessment of environment, health and economic impact of the
676 novel coronavirus (COVID-19). Environ Dev Sustain 2020a: 1-11.
677 Sarkodie SA, Owusu PA. Impact of meteorological factors on COVID-19 pandemic: Evidence
678 from top 20 countries with confirmed cases. Environ Res 2020b; 191: 110101.
679 Sorci G, Faivre B, Morand S. Explaining among-country variation in COVID-19 case fatality
680 rate. Sci Rep 2020; 10: 18909.

U.S. Census Bureau ACS. American Community Survey 1-Year Estimates, Table DP05. Retrieved from <https://data.census.gov/cedsci/table?q=dp05&tid=ACSDP1Y2018.DP05>, 2018.

Wadhera RK, Wadhera P, Gaba P, Figueroa JF, Joynt Maddox KE, Yeh RW, et al. Variation in COVID-19 Hospitalizations and Deaths Across New York City Boroughs. JAMA 2020; 323: 2192-2195.

Wellenius GA, Burger MR, Coull BA, Schwartz J, Suh HH, Koutrakis P, et al. Ambient air pollution and the risk of acute ischemic stroke. Arch Intern Med 2012; 172: 229-34.

Wu X, Nethery RC, Sabath MB, Braun D, Dominici F. Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis. Sci Adv 2020; 6: eabd4049.

Xing YF, Xu YH, Shi MH, Lian YX. The impact of PM2.5 on the human respiratory system. J Thorac Dis 2016; 8: E69-74.

Yancy CW. COVID-19 and African Americans. JAMA 2020; 323: 1891-1892.

Zangari S, Hill DT, Charette AT, Mirowsky JE. Air quality changes in New York City during the COVID-19 pandemic. Sci Total Environ 2020; 742: 140496.

Zhang JJ, Wei Y, Fang Z. Ozone Pollution: A Major Health Hazard Worldwide. Front Immunol 2019; 10: 2518.

Table 1. Publicly available data sources used in this study.

Data	**Source**
Covid-19 cases & deaths	Coronavirus in NY: Cases, maps, charts, and resources (https://www.syracuse.com/coronavirus-ny/)
Population estimates & demographics 2018	US Census Bureau's American Community Survey (ACS) (https://www.census.gov/programs-surveys/acs)
Temperature & air quality index	EPA (http://www.epa.gov/ttn/airs/aqsdatamart)
Nursing homes locations	The Department of Health and Human Services (HHS) (https://www.arcgis.com/home/item.html?id=b3813b2d3a054c378247bf32bcd8d203)
Satellite $PM_{2.5}$ estimates	Air pollution and COVID-19 mortality in the United States, Harvard University (http://github.com/wxwx1993/PM_COVID)

Figure 1

Click here to access/download;Figure;Figure_1.jpg

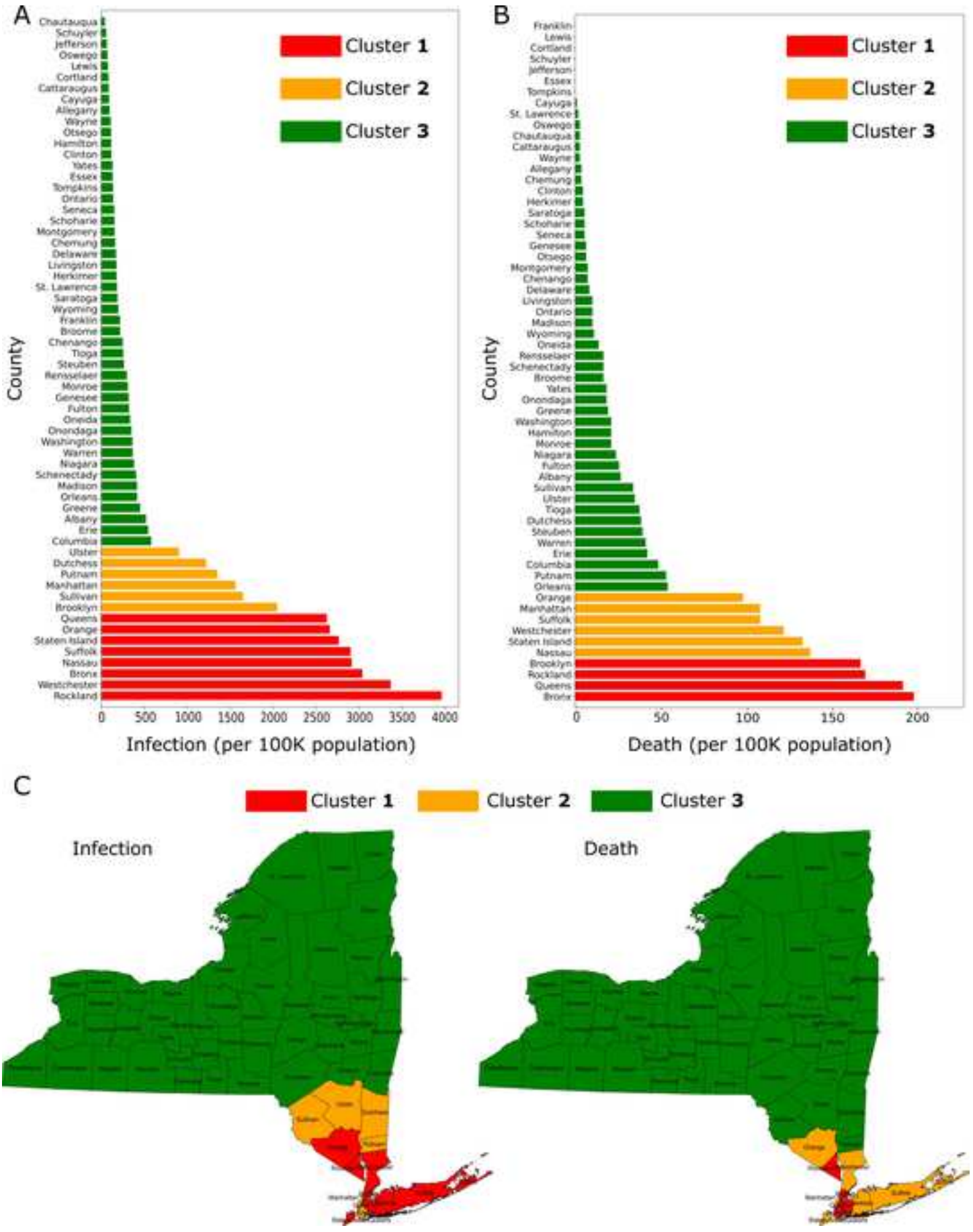


Figure 2

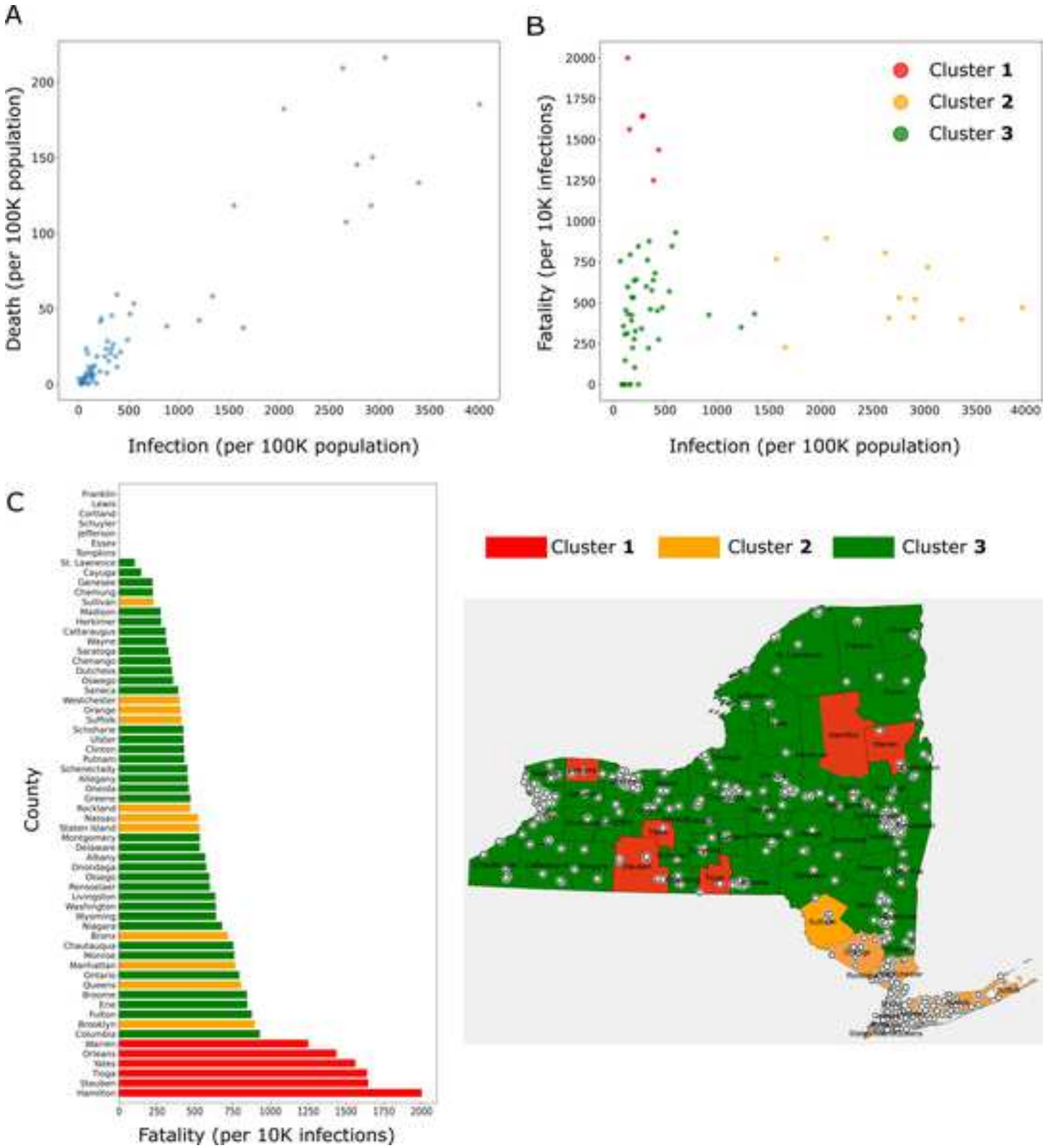


Figure 3

Click here to access/download;Figure;Figure_3.jpg

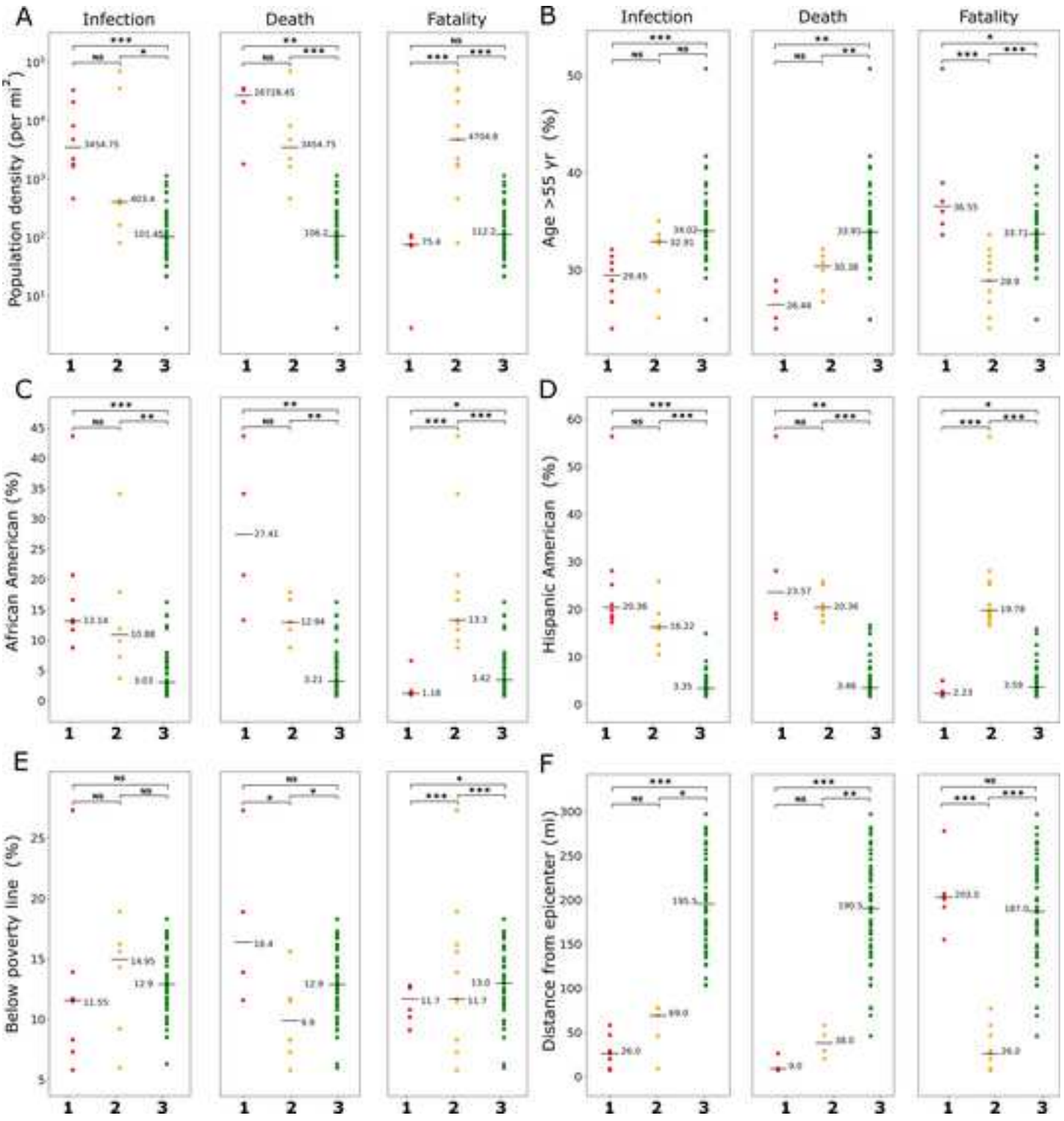


Figure 4

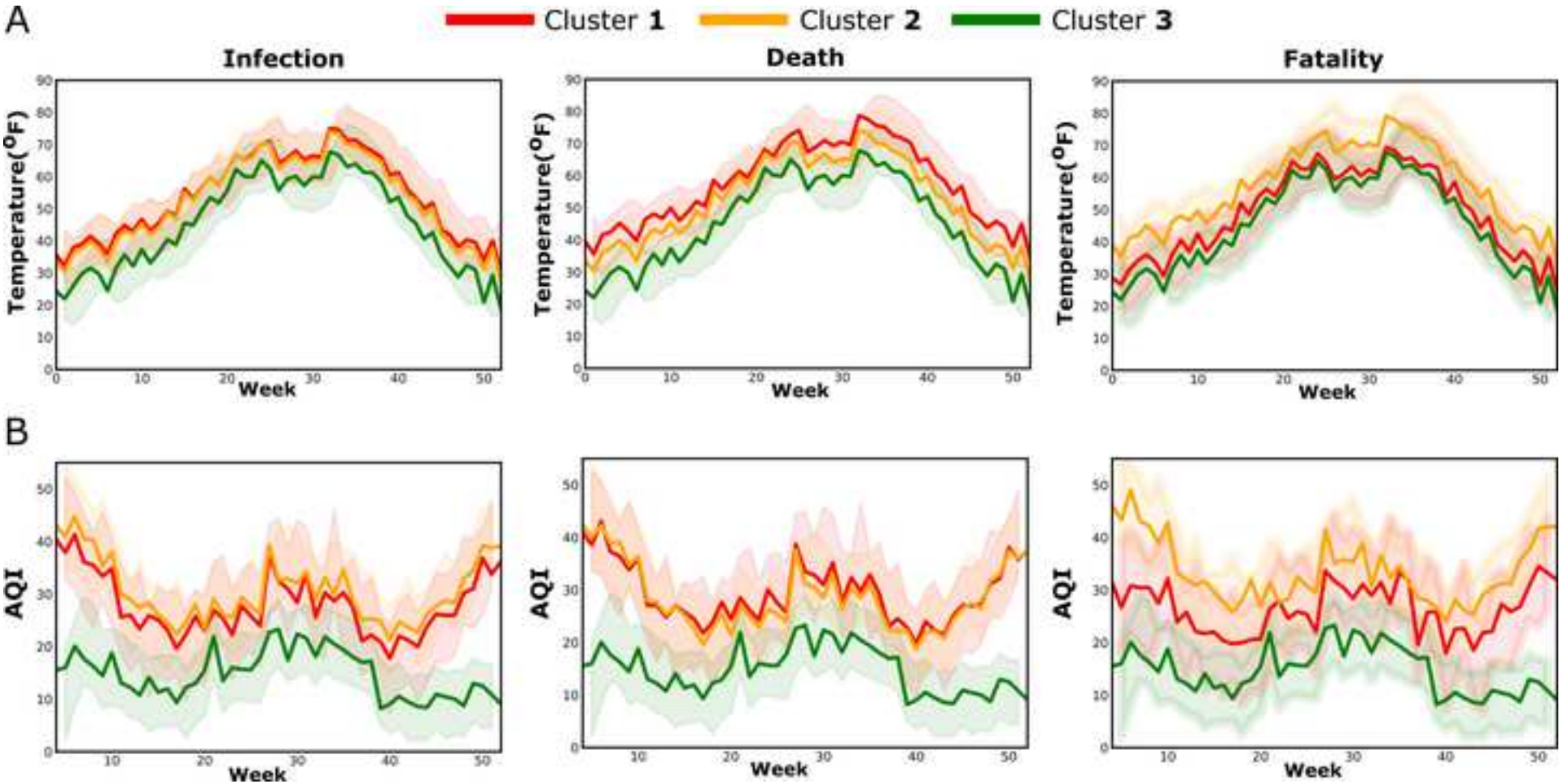


Figure 5

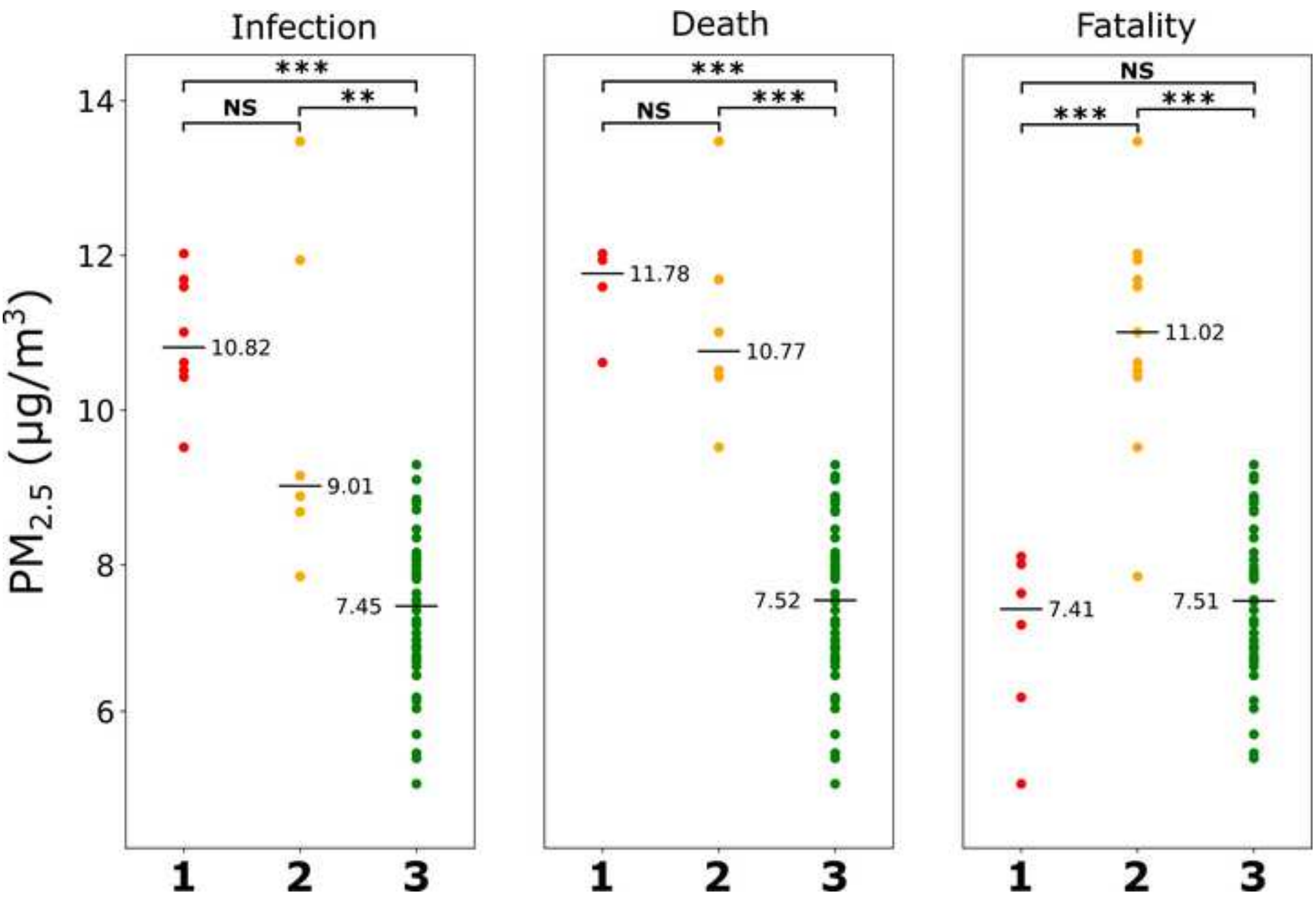
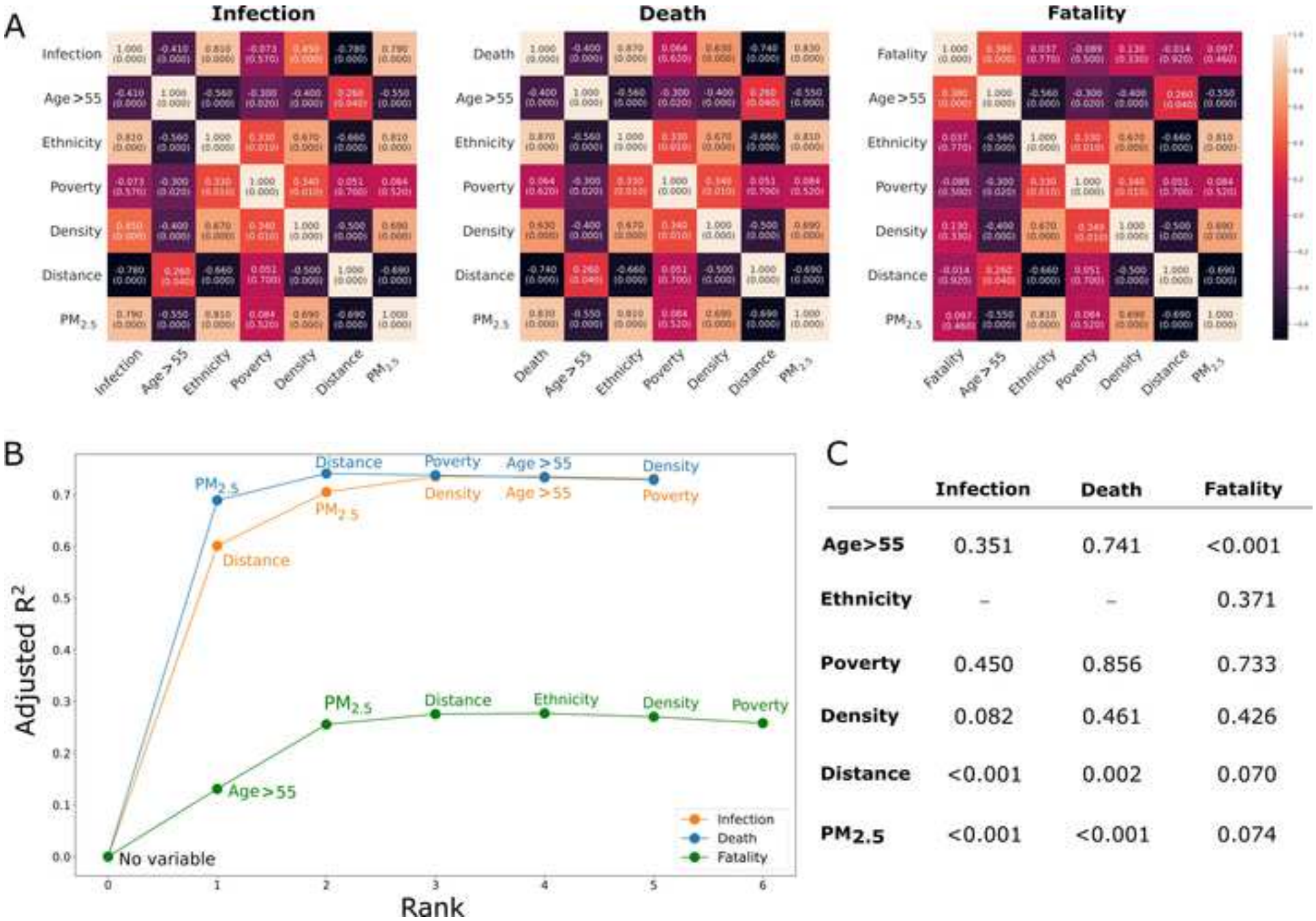


Figure 6

Click here to access/download;Figure;Figure_6.jpg



C

	Infection	Death	Fatality
Age>55	0.351	0.741	<0.001
Ethnicity	–	–	0.371
Poverty	0.450	0.856	0.733
Density	0.082	0.461	0.426
Distance	<0.001	0.002	0.070
PM2.5	<0.001	<0.001	0.074

CRediT authorship contribution statement

Sumona Mondal: Writing – original draft, Supervision, Methodology, Formal analysis, Project administration. **Chaya Chaipitakporn:** Formal analysis, Visualization, Data Curation. **Bridget Wangler**: Visualization. **Vijay Kumar**: Data curation. **Supraja Gurajala**: Review and editing. **Suresh Dhaniyala**: Validation, Review and editing. **Shantanu Sur**: Conceptualization, Investigation, Supervision, Writing - review and editing.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: