Assessing County-Level Risk Factors for COVID-19

Luke Plutowski

5/29/2020

Project Description and Summary

The COVID-19 pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused enormous loss of life and damage to the world economy. Though details about the virus are still emerging, infectious disease experts report that the disease is highly transmittable and quite lethal for those who become infected. Because of uncertainty around the scale and future of the virus, it is necessary to gain a proper understanding of the spread and growth of the disease at an aggregate level to more properly design containment measures. In this paper, we investigate risk factors for COVID-19 across counties in the United States, the nation with the highest reported number of cases and deaths. We apply a variety of statistical learning techniques to study which county-level factors best predict COVID-19 deaths. Our independent variables include a mix of factors related to demographics (e.g. population density), health (e.g. diabetes mortality rates), and politics (e.g. date of stay-at-home orders). Results demonstrate that ...

Literature Review

Research on COVID-19 and its impacts is rapidly evolving and extends into nearly every discipline. Most relevant to this paper are those studies which are using statistical models to forecast the spread of the disease and identify its aggregate-level risk factors. Notably, a research group led by Dr. Bin Yu[1] has undertaken an impressive effort to gather and organize county- and hospital-level data related to the virus from a variety of sources. In a working paper,[2] they test different models for predicting COVID-19 deaths at the county level. They find that an ensemble which combines a linear predictor with 5-day death lags fit separately to each county and and an exponential predictor which is fit to all counties and incorporates information from neighbors performs best. This complements the growing body of literature from other researchers making important contributions to modeling the spread of the epidemic (Elmousalami and Hassanien 2020; Fannelli and Piazza 2020; Killeen et al. 2020; Kucharski et al. 2020; Murray and IMHE 2020).

In this paper, we reanalyze the data from Bin Yu's group, borrowing some parts of their methodology, with a slightly different purpose. Our primary goal is not to exactly predict the number of COVID-19 deaths, but to identify the factors at the county-level that can help us most accurately predict such deaths. This will allow to detect which populations are most vulnerable to the virus and to recommend appropriate mass responses to help reduce mortality. In doing so, we bridge the divide between epidemiological studies which examine individual-level co-morbidities and risk factors associated with coronavirus deaths (Goh et al. 2020; Guan et al. 2020a; Guan et al. 2020b; Wang et al. 2020; Zhou 2020) and studies done by public health experts and researchers which have assessed the impacts of various government interventions and containment measures (Hsiang et al. 2020; Peak et al. 2020; Pei and Shaman 2020). Do counties with more immunocompromised residents experience a greater death toll compared to those with a healthier or younger population? Can a given area's healthcare capacity predict the number of COVID-19 deaths? And do factors related to the political debate surrounding coronavirus, such as the time at which mandatory social distancing orders were adopted, influence how strongly counties are affected by the disease? This paper aims to shed light on these questions.

Following Yu et al., our main outcome of interest is the cumulative number COVID-19 deaths on a given day (April 23). We focus on deaths instead of cases, as the latter is more unreliable due to inconsistent use and availability of testing. We also use the total number of deaths rather than a measure adjusted for

[1] https://www.stat.berkeley.edu/~yugroup/people.html

[2] https://www.stat.berkeley.edu/~binyu/ps/papers2020/covid19_paper.pdf

population as the raw value gives better interpretability of the extent of the virus's penetration and spread into communities across the US.

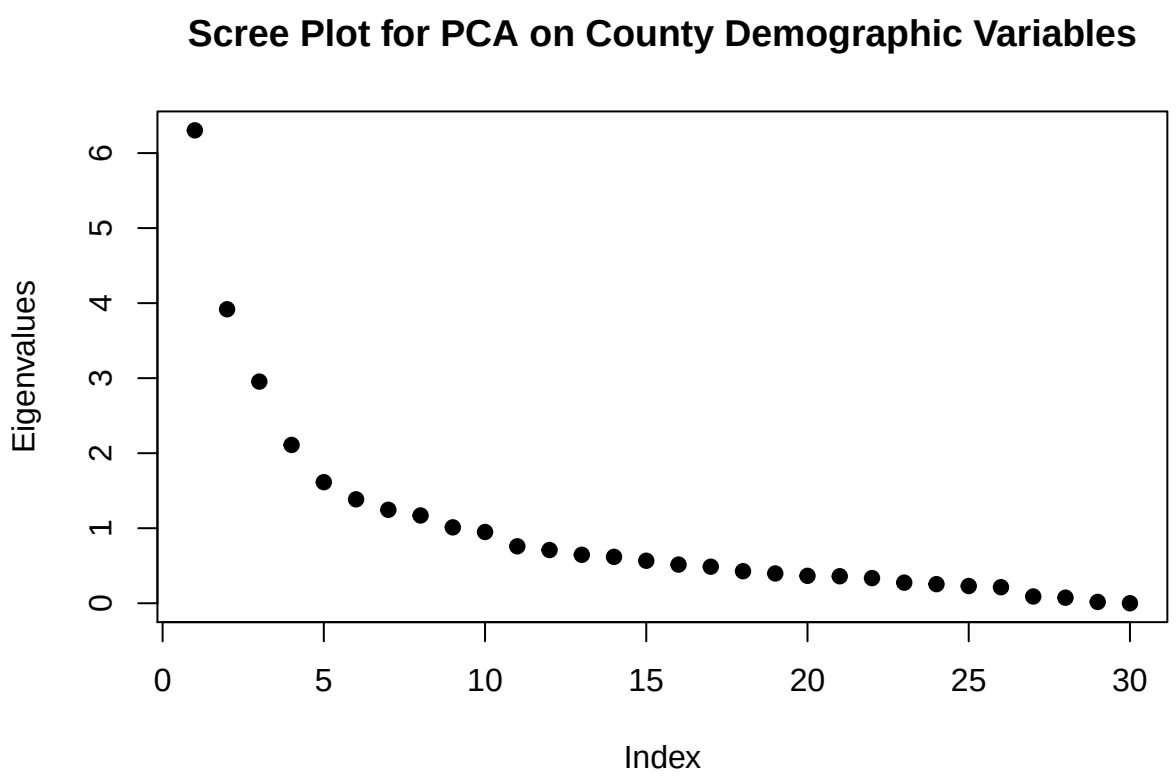
Unsupervised Learning

In this section, we use unsupervised learning methods in order to understand the data and examine relationships between feature variables. For this purpose, we apply three techniques. First, we use principal components analysis (PCA) to reduce county-level demographic features to a few dimensions. Second, we use K-means clustering to group counties based on their COVID-19 case and death counts. Finally, we fuse these two strategies by using the main principal components to cluster counties with hierarchical agglomerative cluster analysis.

Several demographic variables had missing data. We adopted three strategies to address this issue. First, for some variables that had a large proportion (>40%) of values that were missing at random (MAR) (e.g. 3 year diabetes mortality rate or overall mortality rates), we dropped the entire feature from analysis. Second, we imputed values of 0 on some variables for which NAs were meaningful and represented the absence of something in the county, e.g. the shortage of health care professionals or days since a social distancing order was adopted. Finally, for other miscellaneous values that appeared to be missing completely at random (MCAR), e.g. diabetes rate or Democrat-to-Republican ratio, we imputed using the mean within the state.

Principal Components Analysis

Several of the county-level factors seem as though they would be highly correlated with one another. For example, there are different variables related to age within the county, several variables about yearly mortality from various conditions, and multiple variables related to the day in which the county adopted stay-at-home/social distancing orders. To reduce the dimensionality of our feature space, we perform PCA to extract underlying orthogonal principal components from our 30 independent variables.



The independent variables are scaled and centered prior to analysis. From the scree plot above, we see an "elbow" at 5 principal components. Table 1 shows the eigenvalue, proportion of variance explained, and a label/interpretation for each of these components (along with cumulative variance explained by all previous components). The labels are based on the variable loadings. Component 1 loads heavily on age variables (positive values being an older population). Component 2 is highly negative on variables related to disease, smoking, and mortality rates, indicating overall health. The third component has highest loadings on the date at which stay-at-home/social distancing orders were put into place (higher values mean earlier adoption). Component 4 is most closely associated with health care capacity (e.g. health care workers, ICU beds per capita), while component 5, which can be thought of as "overall responsiveness", is negative on both health care capacity and the stay-at-home directives. Together, the five principal components explain 56% of the variation in the 30 independent variables.

Table 1: Summary of Principal Components Analysis for County Features

	PC1	**PC2**	**PC3**	**PC4**	**PC5**
Eigenvalue	2.51	1.98	1.72	1.45	1.27
Prop. Variance	0.21	0.13	0.10	0.07	0.05
Cum. Variance	0.21	0.34	0.44	0.51	0.56
Label	Age	Health	Stay-at-Home	Capacity	Responsiveness

K-Means Clustering

Turning to the variables that capture cumulative daily case and death counts, we use K-means clustering to separate counties into groups based on the severity of their outbreak. For this method, we consider only days within one month of the outcome date (i.e. March 21, 2020 to April 21, 2020). We made this decision

because the sparsity of COVID-19 diagnostic testing in the US makes such counts quite unreliable.[3] Because deaths are necessarily lower than cases, we scale and center each variable to ensure that both types of count variables are used in clustering. However, the overall structure of the clusters does not differ significantly without feature scaling. We use a total of 64 variables for clustering (death and case counts for 32 days).

We try different numbers of clusters (2 through 10) using 25 random initializations. We find that $k = 6$ achieves the best balance of low summed within sum of squares (WSS) and low number of clusters (an "elbow"). Table 2 shows the results of a 6-means clustering algorithm applied to our death/case count variables. For reference, we display the minimum, maximum, and mean of the outcome variable (cumulative number of deaths on April 23) within each cluster, along with the mean number of cases.

Table 2: K-Means Clustering Results for COVID Death/Case Counts

	No. Counties	**Min Deaths**	**Max Deaths**	**Mean Deaths**	**Mean Cases**
Cluster 1	75	19 *(Salt Lake, UT)*	402 *(Hartford, CT)*	106	2405
Cluster 2	19	185 *(Orange, NY)*	849 *(Essex, NJ)*	451	9550
Cluster 3	1	373 *(King, WA)*	373 *(King, WA)*	373	5427
Cluster 5	3	2258 *(Bronx, NY)*	3458 *(Kings, NY)*	3049	37775
Cluster 6	3037	0 *(Union, NM)*	95 *(Hennepin, MN)*	3	66

One cluster, cluster 6,[4] contains the lion's share of counties, all of which had quite low death and case counts. Cluster 3 contained only King County, Washington, which was the early epicenter for COVID-19 in the US. Clusters 1 and 2 captured mid-major metro areas that had small but manageable outbreaks (those in cluster 2 were hit earlier and harder). The remaining two clusters revolved around the most seriously affected counties. Cluster 5 contains the three counties with by far the most cases and deaths, which encompass three of the boroughs of New York City (Brooklyn, Bronx, Queens). Finally, cluster 4 is comprised of four counties surrounding New York City as well as Wayne County, MI (Detroit) and Cook County, IL (Chicago).

We now examine the relationship between the PCA and k-means analyses. Table 3 shows the mean score within each cluster for each of the five principal components, along with mean deaths and mean number deaths per 100 thousand residents. Recalling the PC interpretations, we expect that the hardest hit clusters, i.e. 5, 2, and 3, to be high in PC1 (age) and low in PC2 (health), PC4 (capacity) and PC5 (responsiveness). Component 3 is difficult to predict because stay-at-home orders could be issued in response to high deaths or they may prevent high death counts.

Table 3: K-Means Clusters by Average Principal Component Scores

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**Mean Deaths**	**Mean Deaths/100k**
Cluster 1	-2.86	3.34	0.01	1.12	0.66	106	17
Cluster 2	-3.43	4.12	0.19	1.83	1.01	451	59
Cluster 3	-3.63	6.23	-0.47	1.14	1.13	373	17
Cluster 4	-3.88	6.38	0.25	4.46	2.30	1135	71
Cluster 5	-6.45	7.49	0.00	5.77	4.08	3049	147
Cluster 6	0.11	-0.13	0.00	-0.05	-0.03	3	3

The results do not comport with the expectations; the hardest hit clusters are relatively low in age and high in health, capacity, and responsiveness. This may be occuring because of the imbalance in the number of cases within each cluster and the fact that the counties in clusters 1-5 are so fundamentally different from those in cluster 6 in ways that our principal components are not neatly capturing (e.g. urban vs. rural settings).

[3] Testing continued to be sparse into April and May, and there are certainly self-selection effects, but the counts are not nearly as haphazard and idiosyncratic after mid-March.

[4] the cluster number/names are arbitrary

Hierarchical Clustering

We now tackle the problem in a slightly different way: we first cluster counties based on the principal components using complete linkage hierarchical clustering, then compare those clusters on death and case counts. We choose to use 5 clusters based on examination of a dendrogram and scree plot (not shown).

Table 4 displays the results of this analysis. For each of the five new clusters, we show the average score on the five principal components. For comparison, we also show the mean number of deaths within each cluster. Again, we expect those that score high on PC1 (age) and those that score low on components 2 (health), 4 (capacity), and 5 (responsiveness) to have high death counts.

Table 4: Hierarchical Clusters by Average Principal Component Scores

	No. Counties	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**Mean Deaths**
Cluster 1	11	-6.59	8.46	-0.87	7.96	3.18	1152
Cluster 2	2620	0.41	-0.34	0.05	-0.16	-0.02	5
Cluster 3	466	-2.63	1.53	-0.46	0.50	0.08	32
Cluster 4	40	5.80	1.71	2.23	0.73	0.49	2
Cluster 5	4	-0.76	2.68	-2.08	15.03	-7.66	0

Again, the results are not exactly in line with expectation. Cluster 1 is made up of mostly major metropolitan counties, which are high in deaths and also high in health and health care capacity and low in average age. Clusters 4 and 5 are mostly comprised of rural communities that have major hospitals and high health care capacity and as such have low COVID-19 deaths. Clusters 2 and 3 seem to be catch-alls for all other counties; neither has particularly high COVID death rates nor do they stand out on any components.

Supervised Learning

. . . [continues]. . .

References

- H. H. Elmousalami and A. E. Hassanien. Day level forecasting for Coronavirus disease (COVID-19) spread: Analysis, modeling and recommendations. *arXiv preprint arXiv:2003.07778*, 2020.
- D. Fanelli and F. Piazza. Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos, Solitons & Fractals*, 134:109761, 2020.
- K. J. Goh, S. Kalimuddin, and K. S. Chan. Rapid progression to acute respiratory distress syndrome: Review of current understanding of critical illness from COVID-19 infection. *Annals of the Academy of Medicine, Singapore*, 49(1):1, 2020.
- W. Guan, W. Liang, Y. Zhao, H. Liang, Z. Chen, Y. Li, X. Liu, R. Chen, C. Tang, T. Wang, et al. Comorbidity and its impact on 1590 patients with COVID-19 in China: A nationwide analysis. *European Respiratory Journal*, 2020a.
- W. Guan, Z. Ni, Y. Hu, W. Liang, C. Ou, J. He, L. Liu, H. Shan, C. Lei, D. S. Hui, et al. Clinical characteristics of Coronavirus disease 2019 in China. *New England Journal of Medicine*, 2020b.
- S. Hsiang, D. Allen, S. Annan-Phan, K. Bell, I. Bolliger, T. Chong, H. Druckenmiller, A. Hultgren, L. Y. Huang, E. Krasovich, P. Lau, J. Lee, E. Rolf, J. Tseng, and T. Wu. The effect of large-scale anti-contagion policies on the Coronavirus (COVID-19) pandemic. *medRxiv*, 2020.
- B. D. Killeen, J. Y. Wu, K. Shah, A. Zapaishchykova, P. Nikutta, A. Tamhane, S. Chakraborty, J. Wei, T. Gao, M. Thies, et al. A county-level dataset for informing the united states' response to covid-19. *arXiv preprint arXiv:2004.00756*, 2020.
- A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, , J. Edmunds, S. Funk, and R. M. Eggo. Early dynamics of transmission and control of COVID-19: A mathematical modelling study. *medRxiv*, 2020.
- C. J. Murray and I. H. M. E. COVID-19 health service utilization forecasting team. Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. *medRxiv*, 2020.
- C. M. Peak, R. Kahn, Y. H. Grad, L. M. Childs, R. Li, M. Lipsitch, and C. O. Buckee. Modeling the comparative impact of individual quarantine vs. active monitoring of contacts for the mitigation of COVID-19. *medRxiv*, 2020.
- S. Pei and J. Shaman. Initial simulation of SARS-CoV2 spread and intervention effects in the continental US. *medRxiv*, 2020.
- C. Wang, L. Liu, X. Hao, H. Guo, Q. Wang, J. Huang, N. He, H. Yu, X. Lin, A. Pan, S. Wei, and T. Wu. Evolving epidemiology and impact of non-pharmaceutical interventions on the outbreak of Coronavirus disease 2019 in Wuhan, China. *medRxiv*, 2020.
- F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *The Lancet*, 2020.

Appendix

Data Dictionary

- `tot_deaths` - Response variable: total cumulative COVID-19 deaths as of 4/22/20
- `tot_cases` - Total cumulative COVID-19 cases as of 4/22/20
- `Rural.UrbanContinuumCode2013` - Rural-urban classification continuum (1-9)
- `PopulationEstimate2018` - Population in 2018
- `FracMale2017` - Proportion of population that is male in 2017
- `PopulationDensityperSqMile2010` - Persons per square mile in 2010
- `MedianAge2010` - Median age in 2010
- `Pop_19below`, `Pop_20_44`, `Pop_45_59`, `Pop_60_74`, `Pop_75plus` - percentage of population that falls into age categories: 19 and below, 20-44, 45-59, 60-74, 75+
- `MedicareEnrollment.AgedTot2017` - Proportion of population enrolled in Medicare in 2017
- `HeartDiseaseMortality`, `StrokeMortality`, `RespMortalityRate2014` - Yearly deaths per 100,000 persons from heart disease, stroke, and respiratory failure
- `Smokers_Percentage` - Percentage of population that regularly smokes
- `DiabetesPercentage` - Percentage of population with diabetes
- `X.FTEHospitalTotal2017` - Full time hospital employees per 100k population
- `TotalM.D..s.TotNon.FedandFed2017` - Number of medical doctors per 100k population
- `X.HospParticipatinginNetwork2017` - Number of hospitals participating in network per 100k population
- `X.Hospitals` - Number of hospitals per 100k population
- `ICU_beds` - Number of ICU beds per 100k population
- `SVIPercentile` - In which percentile the county falls on Social Vulnerability Index, a measurement of vulnerability to hazardous events
- `HPSAShortage` - Number of health professionals per 100k population needed to overcome shortage
- `dem_to_rep_ratio` - Ratio of Democratic to Republican voters
- `stay.at.home`, `X.50.gatherings`, `X.500.gatherings`, `public.schools`, `restaurant.dine.in`, `entertainment.gym` - Number of days since 4/22/20 that county adopted social distancing orders: Stay at home, ban on gatherings over 50, ban on gatherings over 500, closing of public schools, closing of dine-in restaurants, closing of entertainment and gym facilities
- `X.Cases_[date]`, `X.Deaths_[date]` - Cumulative number of cases and deaths each day