**Applying machine learning techniques to predict county-level Covid-19 burden with demographic, mobility, and hospital data**

By Erick Andrews, Ninon Becquart, Liu Yuning, and Jenny Wang

**Introduction**

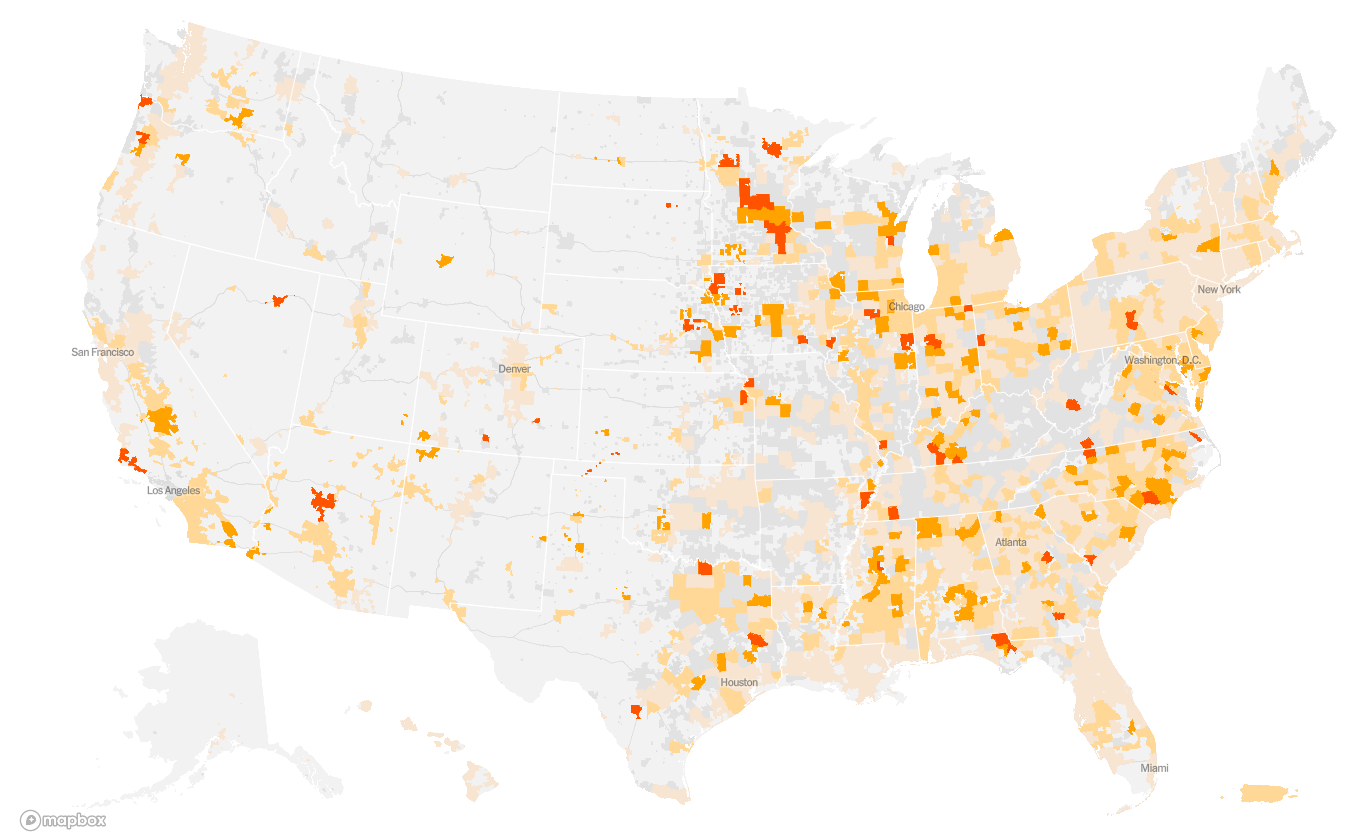
COVID-19 was first introduced to the US on January 15th, when a traveler from Wuhan came back to Seattle, WA, and began showing symptoms of the virus shortly thereafter. Though most thought the first confirmed COVID-19 death occurred in Washington state, new evidence suggests that community spread was occurring weeks prior to the first official death, with one confirmed death in California on February 6th. The failure to contain COVID-19 in the United States has led to an explosion in the number of cases across the country, but particularly in hotspots like New York and now, parts of the Midwest and South. Despite COVID-19 rapidly rising infection rates across large swathes of the country, states have begun to reopen non-essential businesses without sufficient test-and-trace protocol in place (Figures 1 and 2). The IHME, a prominent research institute in Seattle, has revised death and case count predictions through August to the tune of 130,000. The United States has already surpassed their previous estimate of 65,000 deaths from a few weeks ago, which was likely a best case scenario prediction that assumed all social distancing and lockdown measures stayed in place until adequate test-and-trace was implemented (IHME, 2020). Most models, like the one employed by the IHME, incorporate some time-varying aspect to predict future case counts and deaths. Yet the current disease burden may provide as accurate a snapshot as the time-series reliant models, indicating what features of a population have led us to this point. The features that have led to the rapid spread of COVID-10 will likely continue to influence, if not determine, where and when the virus is most severe.

We believe there are several factors that have led to the varied disease burdens by state and county across the US. First, population mobility likely influences disease burden because vectors of the disease come into more frequent contact with the uninfected. Research groups have suggested that the mobility restrictions employed in Wuhan had a major impact on the spread of COVID-19. Indeed, provinces with more introductions of COVID-19 from Wuhan were more likely to have severe outbreaks (Kraemer et al., 2020). Research in the US has found high sensitivity of mobility to stay at home orders and local prevalence of COVID-19 (Engle et al., 2020). Thus, total mobility by county is likely a strong predictor of COVID-19 burden because individuals are COVID-19 vectors, and community spread occurs more quickly without lockdown measures in place.

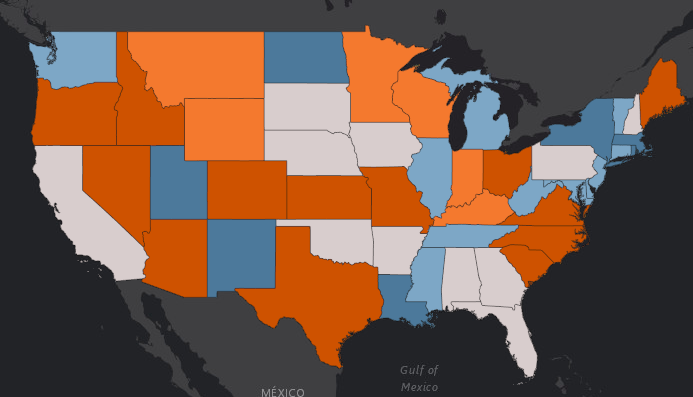
Another factor likely affecting the COVID-19 burden is race. Preliminary research shows that COVID-19 has affected the Hispanic and black communities much more than white communities across the US (Serwer, 2020; AMFAR, 2020). We have also seen that men are disproportionately affected by COVID-19, though the exact reason why remains unknown (NPR, 2020). Marginalized groups in the US are already at elevated risk because of longstanding disparities in the US health system, but the extent to which socioeconomic and racial biases determine COVID-19 disease burden remains understudied and unclear (Laurencin and McClinton, 2020). Adding race to a model that examines disease burden could significantly improve our reopening policies by clarifying the importance of ethnicity in relation to common comorbidities and demographic factors.

Additionally, we suspect that COVID-19 disease burden may be driven by hospital capacity to treat the disease. Many models have focused on predicting hospitalizations and ICU bed needs over the next few months to aid in resource allocation, but few have examined the impact hospital capacity has on COVID-19 burden (Murray, 2020; Remuzzi and Remuzzi, 2020). In our model, inclusion of features representing capacity may be seen as a proxy for the amount of resources invested in a county’s health infrastructure. Finally, we include age as another demographic predictor, hoping to further clarify what age groups are most at risk for further direction of funds and for guiding reopening policy creation.

Many models and research groups have attempted to predict cases and deaths, but not many have focused on the important factors in reopening America. While test-and-trace strategies are undeniably important for reopening, other demographic and social factors could guide policy on a humane, safe approach to reopening the country. Research is scant regarding the burden of COVID-19, as well, because most models seem to focus on time-series analysis for predicting outbreaks and getting ahead of the virus. Now, as states reopen non-essential businesses across the country, research to guide policy on focusing resources in vulnerable, hard-hit areas is of paramount importance.



**Figure 1:** New York Times map of hotspot counties, with orange and red indicating faster rate of cases doubling. Much of the South and Midwest are affected, yet continue to reopen.



**Figure 2:** Missouri University Map of testing rates. Dark blue represents those over 40 per 1,000 residents, and dark orange is less than 18 per 1,000 residents.

**Methods**

We combined county-level data from various sources that contain information on demographics, social mobility, hospital resources, voting preferences, and cumulative COVID-19 case and death counts from January 22 to May 4, 2020. We also calculated the density of cases and deaths due to the varying population across counties.

Race, sex, and age population counts were obtained from the Decennial Census 2010.

Social mobility data was obtained from the global mobility data and the Yu Group at Berkeley.

Voting preferences (Democrat vs. Republican), hospital resource data, and case and death counts were obtained from the Yu Group as well. US population estimates and county FIPS codes as of 2019 were obtained from the US Census Bureau.

The resulting dataset included 1027 observations and over 150 features comprised of:

* Demographics: race, sex, age, and voting preferences
* Social mobility: date lockdown policies were implemented and mobility before vs. after lockdown measures
* Hospital resources: number of hospital beds, number of ICU beds, and average ventilator use
* COVID-19 disease: cases and deaths - total and per capita

*K-means clustering*

We used the k-means algorithm as one of the clustering methods to divide our data. K-means clustering divides data into k clusters, and is an iterative algorithm that functions in the following steps (Kim and Ahn 2008):

1. Initial k centroids are chosen; there are k data points that are far away from each other.
2. Each data point is assigned to the closest centroid, forming a cluster.
3. We adjust the centroid of each cluster to include the new data points assigned to them.
4. Continue iterating through Steps 2 and 3 until clusters no longer change, or stop conditions have been achieved.

We used the k-means function available in the sklearn package from Python. We chose to create 3 clusters based on exploratory analysis, which indicated that less clusters would lump too many dissimilar counties together, but more clusters resulted in certain clusters that would have small sample size, and thus be difficult to effectively analyze. We also found that three clusters best represented the various US counties: rural low population, suburban or medium population, and high population. Counties with missing values were dropped from the analyses.

K-means is a popular choice for clustering algorithms and is an appropriate technique to use in this particular setting due to its ability to perform well in a wide variety of application domains (Wagstaff et al 2001). Multiple studies have shown that despite some of its limitations, such as sensitivity to initial clusteroid selection, K-means has provided good results for practical situations such as anomaly detection and data segmentation (Rodriguez et al 2019). K-means is also an attractive clustering algorithm due to its simplicity and speed compared to more complex clustering algorithms which require higher computational cost.

*Expectation Maximization clustering*

We also used the expectation-maximization (EM) algorithm to cluster the counties. The EM algorithm is an iterative procedure for the maximum likelihood estimate of a parametric distribution underlying some given data (Garriga et al., 2016). We applied the Gaussian Mixture Model that generates a Gaussian distribution of each unknown observation. The EM algorithm functions in the following steps:

1. Initialization: randomly initialize a set of variables .
2. Expectation-step: for each point, compute the likelihood weight, and guess whether the point is more likely to be 0 or 1.
3. Maximization-step: adjust the parameters based on the new assignment to maximize the likelihood of the weights.

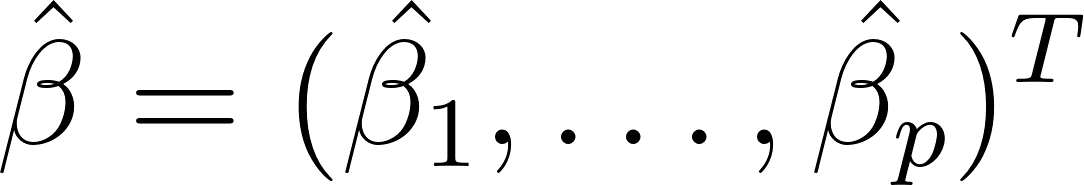
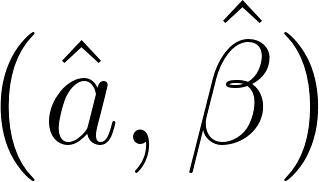
The expectation and maximization steps iterate until convergence is achieved. In the end, each point is assigned to a cluster that it is most probable to fall in. We acknowledge the EM algorithm does not permit us to reach the global maximum of likelihood function, but it guarantees convergence to a local maximization. EM has been documented with a sound performance in previous studies. (McLachlan & Krishnan, 2008; Gupta, Chen, 2010) EM requires a specified number of output clusters. We chose to divide the counties into 3 and 4 clusters, which not only enables us to distinguish different types of counties but also makes the results interpretation and communication easy. We deleted counties with missing values in the analysis.

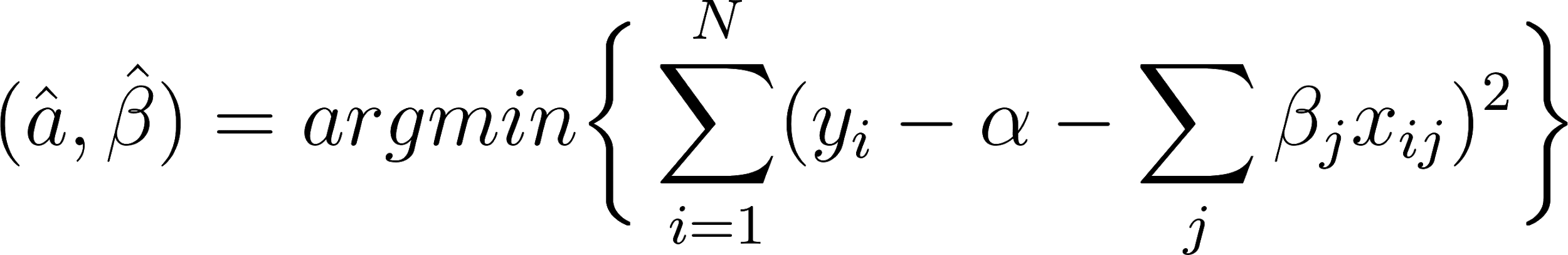
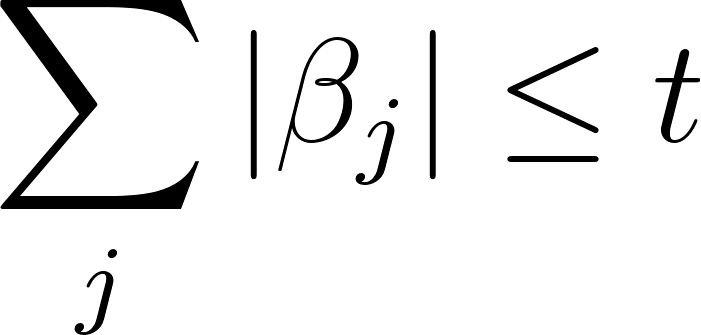
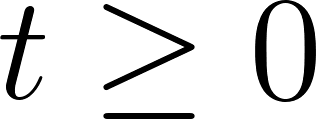
*Predictions*

We used a LASSO model followed by a random forest to predict county-level disease severity using Python’s scikit-learn (Pedregosa et al., 2011). We decided to predict on the log cases per capita as opposed to the raw numbers, because the observations are very close together with one another to model trends using them as-is. We separately ran the models on the observations in each cluster as determined by the clustering algorithms on the assumption that different clusters may have different top predictors associated with their respective disease burdens.

*LASSO*

After clustering, we first used a LASSO model to reduce the dimensionality of the dataset, as we had close to 200 features. Lassos minimize the residual sum of squares contingent on the sum of the absolute value of coefficients equaling less than a constant (Tibshirani 1994). This constraint allows for certain coefficients to be produced that are exactly zero (Tibshirani 1994). Compared to PCA, LASSO gives more interpretable results (Tibshirani 1994). The LASSO equation is adds a penalty term to ordinary least squares:

Letting [](https://www.codecogs.com/eqnedit.php?latex=%5Chat%7B%5Cbeta%7D%20%3D%20(%5Chat%7B%5Cbeta_%7B1%7D%7D%2C%20%5Cdots%2C%20%5Chat%7B%5Cbeta_%7Bp%7D%7D)%5E%7BT%7D#0), the LASSO estimate [](https://www.codecogs.com/eqnedit.php?latex=(%5Chat%7Ba%7D%2C%20%5Chat%7B%5Cbeta%7D)#0) is defined by

[](https://www.codecogs.com/eqnedit.php?latex=(%5Chat%7Ba%7D%2C%20%5Chat%7B%5Cbeta%7D)%20%3D%20argmin%20%5Cbigg%7B%5C%7B%7D%20%5Csum_%7Bi%3D1%7D%5E%7BN%7D%20(y_%7Bi%7D%20-%20%5Calpha%20-%20%5Csum_%7Bj%7D%20%5Cbeta_%7Bj%7D%20x_%7Bij%7D)%5E%7B2%7D%20%5Cbigg%7B%5C%7D%7D#0), subject to [](https://www.codecogs.com/eqnedit.php?latex=%5Csum_%7Bj%7D%20%7C%5Cbeta_%7Bj%7D%7C%20%5Cleq%20t#0). Here, [](https://www.codecogs.com/eqnedit.php?latex=t%20%5Cgeq%200#0) is a tuning parameter.

For each run of the model, we divided the respective cluster dataset into training and test sets, and we tuned the hyperparameters using 5-fold cross validation. The best hyperparameter was chosen as the one that maximized the cross validation score ± standard error. We then ran the model again using this hyperparameter and chose the 50 most important predictors, which were ones with the largest coefficients after the L1 penalty.

*Random forest*

We then subsetted the dataset for each cluster using only the top 10 predictors determined by the LASSO model and ran a random forest with those predictors for each cluster. The random forest is an ensemble method, which generally performs better than any single model alone. This ensures a low correlation across individual trees and favors the prediction rule obtained from the tree with the most votes. Since a single decision tree is high-variance and prone to overfitting, the collection of these uncorrelated trees protect each other from their individual errors. The random forest uses different combinations of predictors in its trees. Each tree draws a random sample from the original training data, and the number of baggings (estimators) was the one that maximized the model fit (R2). Each tree split on 10 predictors as the features in our dataset were not highly correlated. We used varying numbers of trees for each cluster, as they were of different sizes and thus required variable splits and depth. The random forest is a low-variance model that performs with high accuracy in regression settings and is therefore suitable for our goal. However, random forests can be computationally intensive and not very interpretable due to the nature of ensemble methods. Additionally, the random forest can overfit when the features are highly correlated with each other and perform with lower accuracy when there is a lot of noise in the dataset. We have removed predictors that may be correlated with others, such as total population. Thus, the random forest is a suitable model for the purposes of this dataset.

**Results**

Cluster: KM & EM

Shown in Figure 3, we performed K-means and EM for both 3 clusters and 4 clusters. There are 786, 225, and 16 counties in each of the clusters from the K-means algorithm when specifying 3 cluster (Figure 3(A)), and 701, 216, 94, and 16 counties in each cluster from K-means with 4 clusters (Figure 3(C)). The output clusters from the EM algorithm with 3 clusters contained 580, 288, and 159 counties (Figure 3(B)), and the EM algorithm with 4 clusters contained 566, 286, 132, and 43 counties in each cluster (Figure 3(D)).

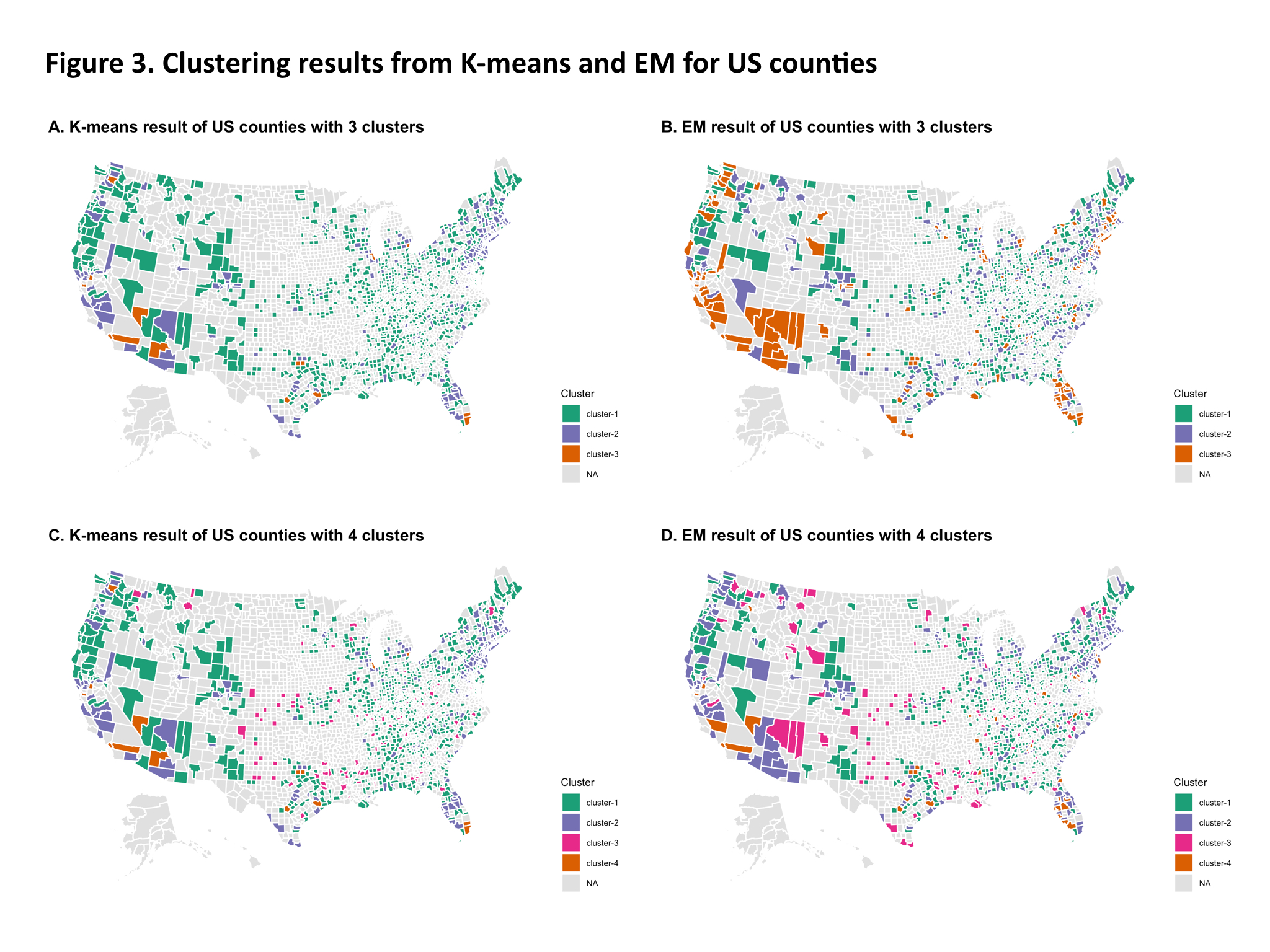
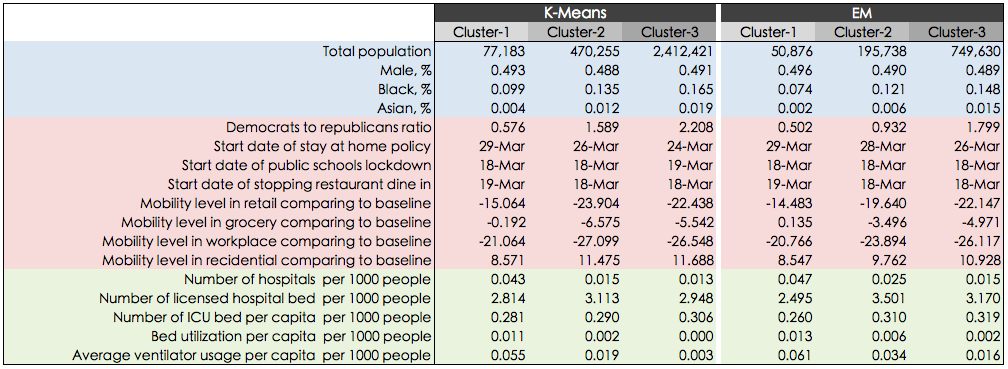


Table 1 presents the characteristics of each cluster from the K-means and EM algorithms. In both algorithms, clusters are presenting substantial divergence in total population, the proportion of black population, and the Democrats to Republicans ratio. Cluster 3 has the largest population, a higher proportion of black population, and more Democrats relative to Republicans. The clusters with a larger population count also tend to have earlier start date of stay-at-home policy and more drop in the mobility level in retail, grocery, and workplace. No significant difference in the hospital resource, presented by hospital and hospital beds per capita, are found among clusters. However, counties in the cluster with larger population size have fewer bed utilization and average ventilator usage per capita than those with smaller population size, indicating a potential crisis of hospital resources under COVID-19 in the urban areas. We choose to use the result from K-means to continue the following study.

**Table 1**. The characteristics of clusters from K-means and EM.



**Table 2.** Most important 10 features determined by lasso for each cluster.

|  |  |
| --- | --- |
| *Cluster* | *Features* |
| Cluster 1 | * Date stay at home order was implemented * Date gatherings of 50+ were stopped * Pacific Islander proportion of population * Date entertainment/gym spaces were closed * Date gatherings of 500+ were stopped * Date public schools were closed * Date restaurants were closed (dine-in) * Total Male population * Total Female population * Total Monoracial Population |
| Cluster 2 | * Other Race proportion of population * Pacific Islander proportion of population * Democrat to Republican ratio * Date stay at home order was implemented * Date gatherings of 50+ were stopped * Date gatherings of 500+ were stopped * Date public schools were closed * Date restaurants were closed (dine-in) * Date entertainment/gym spaces were closed * Total Male population |
| Cluster 3 | * 65-74 year olds proportion of population * Date gatherings of 500+ were stopped * Date entertainment/gym spaces were closed * Mobility- parks * Number of hospitals * Mobility- grocery stores * Potential increase in bed capacity * Average ventilator usage * Date stay at home order was implemented * Total Male population |

**Table 3.** Top 5 important predictors determined by Random Forest.

|  |  |  |
| --- | --- | --- |
| *Cluster* | *Features* | *Importance* |
| Cluster 1 | Date of entertainment/gym closures | 0.51 |
| Date of stay at home order | 0.23 |
| Pacific Islander proportion of population | 0.14 |
| Date of 50+ gatherings order | 0.06 |
| Monoracial proportion of population | 0.03 |
| Cluster 2 | Date of 50+ gatherings order | 0.67 |
| Democrats to Republicans ratio | 0.17 |
| Date of stay at home order | 0.09 |
| Date of 500+ gatherings order | 0.03 |
| Date of entertainment/gym closures | 0.02 |
| Cluster 3 | Mobility - park areas | 0.38 |
| Number of hospitals | 0.38 |
| Date of entertainment/gym closures | 0.07 |
| Potential increase in bed capacity | 0.07 |
| Date of 500+ gatherings order | 0.04 |

Supervised learning:

*Lasso*

We show the top 10 predictors (out of 50) determined by LASSO for each cluster in Table 2. Across the clusters, we see several shared variables related to social distancing policies, including the closing of public spaces and large gathering bans. We also see clusters sharing demographic variables, such as the proportion of Pacific Islanders and the total male population. As for predictors related to hospital capacity, these were mostly seen in the urban cluster (cluster 3) and less so for the rural and suburban clusters (clusters 1 and 2, respectively).

*Random Forest*

Using random forest allowed us to see the top predictors for each cluster. One feature showed up in every cluster-- the date of entertainment and gym closures. The importance of this feature varied according to cluster; it is the top important predictor in cluster 1, but only top 5 for cluster 2, with significantly different importance measures. However, each cluster contained unique top predictors that were not represented in other clusters. Pacific Islander proportion of population was quite important in Cluster 1, Democrat to Republican ratio was very important in Cluster 2 but nowhere else, and in Cluster 3, mobility in parks (a measure of social distance), was of great importance.

**Discussion**

While the majority of research on COVID-19 has focused on case counts, death counts, and prediction, we chose instead to investigate burden on the county-level, in order to understand resource distribution, visualize disparate impact across the U.S., learn which predictors were most important-- all with the hopes of providing more information for reopening strategies.

From our cluster analyses, we saw three or four clusters of counties, sometimes with quite varied characteristics. Population size is one of the most distinguishing variables of the clusters. The cluster with bigger population size is characterized by higher proportions of Black and Asian population, more Democrats, earlier stay at home policy, more bed utilization and ventilator usage per capita. The cluster with middle population size has the most decrease in mobility level. Fewer proportion of Black and Asian population, more Republicans, and the least decline in mobility characterized the cluster with the smallest population size. Both KM and EM generalized the three clusters of “mega-city”, “middle-size city”, and “township”. We found the hospital resources, such as hospital bed per capita or ICU bed per capita, were in the similar level in all clusters. However, the hospital bed and ventilator were used more intensively in the mega-city cluster, indicating a more severe outbreak in the mega-city clusters. This result implies the need to consider county-specific contexts when designing reopening policy. More healthcare service provisions are crucial for the counties in the mega-city cluster. The impact of stopping lockdown and mobility resurgence should be a major concern for counties in the middle-size city cluster.

From our lasso analysis, we were able to filter out less important features from the 3 clusters identified by our k-means clustering. From Table 2, we see that on a broad level, the main predictors from the three clusters are more or less similar. This indicates that there are common variables associated with disease regardless of whether a region is urban, suburban, or rural. These may be nation-wide trends that deserve more scrutiny, which we cover in the random forest.

From our Random Forest, we were able identify top predictors in each of our 3 clusters. The appearance of ‘Date of closure of entertainment/gyms’ in every cluster illustrates the possible nationwide importance of this variable regarding disease burden. This is something that all government and health officials could consider when designing reopening strategies and deciding which businesses to open at what phase. Each cluster also contained top predictors that were unique to it; this confirms our suspicion that counties may have different needs and priorities to evaluate. In Cluster 3, mobility in parks was the number one important predictor as selected by Random Forest, despite this feature not appearing in any other cluster. Therefore, officials in such counties could think about how to minimize public interactions in parks, or encourage their constituents to spread out perhaps.

**Conclusion**

The demographic, mobility, hospital capacity, and social distancing measures taken after the emergence of COVID-19 is not uniform across the United States. We divided the population into 3 clusters and found that individual counties have disparate disease burdens with different factors to consider. At the state level, governors and other executive branches of government usually make decisions that apply to the entire state. However, they may want to consider making decisions on a county-level basis, since counties vary in their COVID-associated features across the state. Reopening the nation may be sooner and more effective by prioritizing the allocation of resources to populations that need them the most.

**References**

AMFAR. (2020). Ending the HIV Epidemic Database. https://ehe.amfar.org/inequity?\_ga=2.51214761.1618924293.1588715818-1730120696.1588715818

Buckee, C. O., Balsari, S., Chan, J., Crosas, M., Dominici, F., Gasser, U., Grad, Y. H., Grenfell, B., Halloran, M. E., Kraemer, M. U. G., Lipsitch, M., Metcalf, C. J. E., Meyers, L. A., Perkins, T. A., Santillana, M., Scarpino, S. V., Viboud, C., Wesolowski, A., & Schroeder, A. (2020). Aggregated mobility data could help fight COVID-19. Science, 368(6487), 145–146. https://doi.org/10.1126/science.abb8021

Engle, S., Stromme, J., & Zhou, A. (2020). Staying at Home: Mobility Effects of COVID-19 (SSRN Scholarly Paper ID 3565703). Social Science Research Network. <https://doi.org/10.2139/ssrn.3565703>

Garriga, J., Palmer, J. R., Oltra, A., & Bartumeus, F. (2016). Expectation-maximization binary clustering for behavioural annotation. PLoS One, 11(3).

Gupta M, Chen Y. (2010) Theory and use of the EM Algorithm. Foundations and Trends in Signal Processing. 2010; 4(3):223–296. doi: 10.1561/2000000034

IHME. (2020). COVID-19 Projections. Institute for Health Metrics and Evaluation. https://covid19.healthdata.org/

Kraemer, M. U. G., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D. M., Group†, O. C.-19 D. W., Plessis, L. du, Faria, N. R., Li, R., Hanage, W. P., Brownstein, J. S., Layan, M., Vespignani, A., Tian, H., Dye, C., Pybus, O. G., & Scarpino, S. V. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. Science, 368(6490), 493–497. <https://doi.org/10.1126/science.abb4218>

Kim, K & Ahn H (2008). A recommender system using GA k-means clustering inn an online shopping market, 34(2008): 1200-1209. <https://doi.org/10.1016/j.eswa.2006.12.025>

Laurencin, C. T., & McClinton, A. (2020). The COVID-19 Pandemic: A Call to Action to Identify and Address Racial and Ethnic Disparities. Journal of Racial and Ethnic Health Disparities. <https://doi.org/10.1007/s40615-020-00756-0>

McLachlan G, Krishnan T. (2008) The EM algorithm and extensions. 2nd ed. Wiley series in probability and statistics. Hoboken, NJ: Wiley; 2008.

NPR. (2020, April 23). Researchers Study Why Men Seem To Be More Affected By COVID-19. NPR.Org. <https://www.npr.org/2020/04/23/842195564/researchers-study-why-men-seem-to-be-more-affected-by-covid-19>

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Remuzzi, A., & Remuzzi, G. (2020). COVID-19 and Italy: What next? The Lancet, 395(10231), 1225–1228. <https://doi.org/10.1016/S0140-6736(20)30627-9>

Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa LdF, et al. (2019) Clustering algorithms: A comparative approach. PLoS ONE 14(1): e0210236. <https://doi.org/10.1371/journal.pone.0210236>

Serwer, A. (2020, May 8). The Coronavirus Was an Emergency Until Trump Found Out Who Was Dying. The Atlantic. https://www.theatlantic.com/ideas/archive/2020/05/americas-racial-contract-showing/611389/

Team, I. C.-19 health service utilization forecasting, & Murray, C. J. (2020). Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. MedRxiv, 2020.03.27.20043752. <https://doi.org/10.1101/2020.03.27.20043752>

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. J. R. Statist. Soc. B 58(1): 267-288.

Times, T. N. Y. (2020). Coronavirus in the U.S.: Latest Map and Case Count. The New York Times. https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html

Yancy, C. W. (2020). COVID-19 and African Americans. JAMA. https://doi.org/10.1001/jama.2020.6548