Memo: Replicating County Clustering in Wallace, Sharfstein, and Kaminsky (2019)

From: Eric LaRose
To: Professor Brodnax

Project Proposal:

For my project I plan to replicate a paper titled "Comparison of US County-Level Public Health Performance Rankings with County Cluster and National Rankings," appearing in the *JAMA Network Open* journal (Wallace, Sharfstein, and Kaminsky, 2019). The paper uses county-level data on economic and demographic characteristics from the Census Bureau's American Community Survey ("American Community Survey Data," 2020), as well as county-level data on health characteristics from the Robert Wood Johnson Foundation ("Explore Health Rankings," 2020), in addition to individual-level data on these characteristics from the CDC's Behavioral Risk Factor Surveillance System Survey ("Behavioral Risk Factor Surveillance System," 2020). The authors first use a "random forest algorithm to rank the importance of variables present in the individual-level data" in relation to individual health outcomes (Wallace, Sharfstein, and Kaminsky, 2019), and then use the county-level versions of these corresponding variables. Using this county-level data, they then "used k-means analysis to identify [eight] clusters of counties with similar sociodemographic profiles" (Wallace, Sharfstein, and Kaminsky, 2019). They then assign percentile ranks to counties on health outcomes both among counties nationwide and among counties within their cluster. These cluster-based rankings help improve on nationwide rankings by comparing counties to other counties that have relatively similar demographic characteristics.

My replication will differ from, and expand on, the paper in the following ways. First, while the authors use random forests as a dimension reduction technique, I plan to consider other dimension reduction techniques such as principal components analysis, as detailed in James, Witten, Hastie, and Tibshirani (2017), and see whether different techniques produce similar sets of "important" variables. Next, the authors use a k-means approach to cluster counties with $k = 8$, but provide no justification on the choice of k. I plan to first explore many different potential values of k to see which value provides the lowest SSE, and compare cluster results using this value and the value of $k = 8$ (assuming they are not the same). Additionally, I plan to use other clustering techniques such as hierarchical clustering and density-based scanning, as described in Tan, Steinbach, Karpatne, and Kumar (2019), to see whether these different clustering methods produce clusters that are generally similar. One additional minor difference is that the authors generally use data from the period 2010 through 2014, while their variables are now generally available for 2018 and 2019. I plan to use both the authors' data and the most recent data available, to see whether changes in county characteristics over the past roughly half-decade have substantially changed the results of this analysis. Additionally, I tend to incorporate a spatial component into my analysis to see if rough regional labels can be applied to this analysis. (For instance, one cluster of counties may be largely described as "rural Southern counties.") Such labels will make it much easier to communicate my findings to a non-technical audience. Finally, I will compare my results to those of the American Communities Project, which uses clustering techniques to group counties into 15 clusters based on nearly 40 demographic characteristics ("American Communities Project," n.d.). Because my project will involve dimension reduction, this comparison will provide an idea of the extent to which dimension reduction may be affecting the results.

Works Cited

American Communities Project. American Communities Project. Retrieved October 7, 2020, from https://www.americancommunities.org/

Centers for Disease Control and Prevention. (2020, August 31). Behavioral Risk Factor Surveillance System. Retrieved October 7, 2020, from https://www.cdc.gov/brfss/index.html

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: with applications in R*. New York: Springer.

Robert Wood Johnson Foundation. (2020). Explore Health Rankings. Retrieved October 7, 2020, from https://www.countyhealthrankings.org/explore-health-rankings

Tan, P., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to data mining*. New York: Pearson Education.

US Census Bureau. (2020, March 20). American Community Survey Data. Retrieved October 7, 2020, from https://www.census.gov/programs-surveys/acs/data.html

Wallace, M., Sharfstein, J., & Kaminsky, J. (2019). Comparison of US County-Level Public Health Performance Rankings With County Cluster and National Rankings. *JAMA Network Open*, *2(1)*. doi: 10.1001/jamanetworkopen.2018.6816