Clustering Analysis of COVID-19 Responses by County to Inform Future Pandemic Prevention
The Eastcoast Eagles
Andrew Faris, Jyothi Karra, Morteza Maleki, Ruchi Patel, Betsy Thorne, Tony Bakshi

Abstract:

(Q1) COVID-19 has had devastating impacts on the United States (US) population. The pandemic impacted certain populations differently than others based on demographics [3,7,8,9], financial status [3,18], behavioral / psychographics [1,4,6,7,16] and geographies [1,9]. Additionally, socio-economic factors like exposure to media, political party affiliation[2], etc. played an important role in influencing vaccine uptake [4] and different jurisdictions (states, counties, federal, etc.) had varying responses on how to control and reduce risk of the pandemic (e.g., mask mandates, school closures) [1,5]. Our objective is to identify "high-performing" groups of counties (i.e., counties that experienced below-average COVID impact), execute summary statistics about why the groups might be "high-performing", and analyze "lower-performing" (i.e., counties that experience high COVID impact) groups of counties to suggest improvement opportunities. With this analysis, counties can use data-driven insights to instill new practices into how to manage COVID-19.

Our approach:

(Q2) Currently, the United States is following a broad, conversative approach to contain the pandemic that focuses strongly on preventing transmission[1]. This is logical given the contagious nature of the virus but there might be underlying systemic differences putting certain groups of society at a bigger disadvantage than others [5,9]. Hesitancy to vaccines has been cited as a barrier to effective control of COVID-19[6] but we need to understand the root-cause of this hesitancy to better handle future unforeseen circumstances. Previous analysis on the spread of the pandemic that is geography-specific is mostly limited to the number of "cases", "deaths" and "vaccines"[11,12,17,21]. This is due to the availability of data broadly[14] and the limitations of those data that do exist [9,13,19]. While current approaches are informative in nature, they lack the prescriptive aspect of data analysis and fail to provide recommendations as to what actions, if any, should be taken to promote vaccination.

(Q3) In our approach, an unsupervised learning analysis of COVID-19 infection, death rate, and many other population factors by county will be carried out to identify factors that lead to the specific COVID-19 outcomes in the county. Counties will be clustered into like-groupings (e.g., k-clusters), and the groupings will be analyzed for key summary statistics like average hospitalization rate, average vaccine adoption, average education status, etc. By doing so, assumptions can then be drawn from the summary statistics as to what attributes to a "higher performing" county vs. a "lower performing" county. The groupings can then be visually analyzed by utilizing the first two principal components to identify the separation and distance from other groupings.

While K-Means has been used in some existing studies, this approach combines the algorithm with the demographic and socio-economic data [2,3,4,5,6,7,8,9] by county and provides a fresh perspective on the drivers of pandemic spread to identify communities that are likely to respond well and poorly to unforeseen future events[20] and help identify pain points that need to be worked on at the policy level to help communities be better prepared for a similar health emergency in the future.

(Q4) This study will help inform future research on ways to evaluate the effectiveness of public health preparedness, gauge variables impacting people's response to governmental mandates, and influencing factors in containing a pandemic of similar and bring to light any county-specific factors that can be targeted and corrected using effective public policy measures.

The case for this work (cost, tradeoffs, and benefits):

(Q5) The team's goal is for the results to provide communities with a concise dataset that shows examples of counties that performed well throughout the pandemic, therefore providing a roadmap for improved responses to future pandemics. This, combined with a follow-up rigorous analysis of current and past policy measures, can potentially save lives, reduce strain on hospital systems, and mitigate economic impact. Measurement of success will be based on how each community continues to report out impact from COVID-19. If our identified groupings of "high-performing" communicates continue to report better results than the groupings of "low-performing" communities, then one can argue that the analysis is a success and next steps should be taken.

(Q6) Using demographic data in the wrong context can lead to biased analysis and lead to improper recommendations for policy making. The analysis needs to ensure that we are identifying or eliminating any biases when comparing data across counties – our clustering approach is specifically meant to reduce unrealistic or inhumane policy recommendations while providing data-driven context on COVID-era performance. On completion, the study should identify influencing factors and inform future studies to improve human lives[16], reduce economic impact on society and strain on the healthcare system.

(Q7) The monetary cost of the initial study is minimal; most data is publicly available[13]. The algorithms that will be utilized in the paper will need basic computational power. There is a significant opportunity cost involved in terms of time. We are estimating this project will take ~20 hours a week for 8 weeks, which includes: data aggregation, model building, analysis and report writing.

(Q8) The initial study will require ~2 months of iterative analysis and insight documentation. This is considering cooperation among six analysts. Further exploration and advancement of the study will take more time and study; potentially 1-2 years.

Monitoring Progress:

(Q9) For the initial study, we can identify midterm and final "exams" to check for success. Our midterm check for success will be based on the ability to identify logical groupings of counties based on the available data. If our clustering algorithm can identify reasonable groupings of the counties based on the data, and the summary statistics about those groupings are reasonable, then we can confirm that the study is on the right track for success. Our final check for success will be based on the ability to make reasonable / actionable recommendations to counties to improve their response to COVID-19 and future pandemics. Additionally, the final check for success will be based on the ability to outline clear next steps for how the study can be continued for more robust hypothesis testing.

Work breakdown structure:

Name	Task(s)	Start Date	Target date
Andrew Faris	Model Dev, Report Writing, Poster Building	1/25/2022	4/23/2022
Jyothi Karra	Data Engineering, Report Writing, Poster Building	1/25/2022	4/23/2022
Morteza Maleki	Data Engineering, Report Writing, Poster Building	1/25/2022	4/23/2022
Ruchi Patel	Model Dev, Report Writing, Poster Building	1/25/2022	4/23/2022
Betsy Thorne	Visualization, Report Writing, Poster Building	1/25/2022	4/23/2022
Tony Bakshi	Model Dev, Report Writing, Poster Building	1/25/2022	4/23/2022

All team members have contributed a similar amount of effort.

Bibliography:

1. Interdependence and the cost of uncoordinated responses to COVID-19; Authors: Holtz, David and Zhao, Michael and Benzell, Seth G. and Cao, Cathy Y. and Rahimian, Mohammad Amin and Yang, Jeremy and Allen, Jennifer and Collis, Avinash and Moehring, Alex and Sowrirajan, Tara and Ghosh, Dipayan and Zhang, Yunhao and Dhillon, Paramveer S. and Nicolaides, Christos and Eckles, Dean and Aral, Sinan; http://www.pnas.org/content/117/33/19837.abstract;
2. Political Beliefs affect Compliance with Government Mandates; Authors: Marcus Painter, Tian Qiu https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3569098 ,
3. The relationship between vaccination rates and COVID-19 cases and deaths in the USA Authors: Ensheng Dong and Lauren Gardner; https://systems.jhu.edu/research/public-health/covid-19-vaccine/?fbclid=IwAR21qH7AMfCjDS-PE1BTJyQO4vodMUZhWAdkmHp5v653hDSgZFTqusrKAEY ;
4. Individual and social determinants of COVID-19 vaccine uptake; Authors: Viswanath, K., Bekalu, M., Dhawan, D. et al. Individual and social determinants of COVID-19 vaccine uptake. BMC Public Health 21, 818 (2021) https://doi.org/10.1186/s12889-021-10862-1 ;
5. Racial, Economic, and Health Inequality and COVID-19 Infection in the United States; Authors: Abedi, V., Olulana, O., Avula, V., Chaudhary, D., Khan, A., Shahjouei, S., Li, J., & Zand, R. (2021). *Journal of racial and ethnic health disparities*, *8*(3), 732–742. https://doi.org/10.1007/s40615-020-00833-4
6. Hesitancy in the time of coronavirus: Temporal, spatial, and sociodemographic variations in COVID-19 vaccine hesitancy; Authors: Ran Liu, Gabriel Miao Li, SSM - Population Health, Volume 15,2021,100896,ISSN 2352-8273; https://doi.org/10.1016/j.ssmph.2021.100896 ;
7. Predictors of willingness to get a COVID-19 vaccine in the U.S; Authors: Kelly, Bridget J., Southwell, Brian G., McCormack, Lauren A., Bann, Carla M., MacDonald, Pia D. M., Frasier, Alicia M., Bevc, Christine A., Brewer, Noel T., Squiers, Linda B. https://doi.org/10.1186/s12879-021-06023-9 DO: 10.1186/s12879-021-06023-9
8. Covid-19 by Race and Ethnicity: A National Cohort Study of 6 Million United States Veterans; Authors: Rentsch, Christopher T. and Kidwai-Khan, Farah and Tate, Janet P. and Park, Lesley S. and King, Joseph T. and Skanderson, Melissa and Hauser, Ronald G. and Schultze, Anna and Jarvis, Christopher I. and Holodniy, Mark and Lo Re, Vincent and Akg{\"u}n, Kathleen M. and Crothers, Kristina and Taddei, Tamar H. and Freiberg, Matthew S. and Justice, Amy C. https://www.medrxiv.org/content/early/2020/05/18/2020.05.12.20099135
9. The synchronicity of COVID-19 disparities: Statewide epidemiologic trends in SARS-CoV-2 morbidity, hospitalization, and mortality among racial minorities and in rural America Authors: Rentsch, Christopher T. and Kidwai-Khan, Farah and Tate, Janet P. and Park, Lesley S. and King, Joseph T. and Skanderson, Melissa and Hauser, Ronald G. and Schultze, Anna and Jarvis, Christopher I. and Holodniy, Mark and Lo Re, Vincent and Akg, Kathleen M. and Crothers, Kristina and Taddei, Tamar H. and Freiberg, Matthew S. and Justice, Amy C.; https://www.medrxiv.org/content/early/2020/05/18/2020.05.12.20099135 ;
10. Epidemiological and economic impact of COVID-19 in the US; Authors- Chen, J., Vullikanti, A., Santos, J.*et al. Sci Rep***11,**20451 (2021);. https://www.nature.com/articles/s41598-021-99712-z
11. Prediction of the number of COVID-19 confirmed cases based on K-means-LSTM; Authors: Shashank Reddy Vadyala, Sai Nethra Betgeri, Eric A. Sherer, Amod Amritphale, , Array,Volume11,2021,100085,ISSN 2590-0056, https://doi.org/10.1016/j.array.2021.100085 (https://www.sciencedirect.com/science/article/pii/S2590005621000333)
12. COVID-19 Cases and Deaths in Southeast Asia Clustering using K-Means Algorithm: Authors: JuniarHutagalungand Ni Luh WiwikSri Rahayu Ginantraand Gita Widi Bhawikaand Wayan Gede

Suka Parwitaand AnjarWanto and Pawer DarasaPanjaitan; https://iopscience.iop.org/article/10.1088/1742-6596/1783/1/012027
13. Use of Available Data To Inform The COVID-19 Outbreak in South Africa: A Case Study; Authors: Vukosi Marivate, Herkulaas MvE Combrink; https://doi.org/10.48550/arXiv.2004.04813
14. Can auxiliary indicators improve COVID-19 forecasting and hotspot prediction; Authors: DanielJ.McDonald, JacobBien, AldenGreen, AddisonJ.Hu, NatDeFries, SangwonHyun, NataliaL. Oliveira, JamesSharpnack, JingjingTang, RobertTibshirani, Valérie Ventura, LarryWasserman, Ryan J.Tibshirani, Proceedings of the National Academy of SciencesDec 2021,118(51)e2111453118; https://doi.org/10.1073/PNAS.2111453118
15. COVID-19: Short term prediction model using daily incidence data; Authors: Zhao H, Merchant NN, McNulty A, Radcliff TA, Cote MJ, Fischer RSB, et al. (2021). PLoSONE 16(4): e0250110. https://doi.org/10.1371/journal.pone.0250110
16. Socioeconomic status and well-being during COVID-19: A resource-based examination. Journal of Applied Psychology; Authors: Wanberg, C. R., Csillag, B., Douglass, R. P., Zhou, L., & Pollard, M. S. (2020)., 105(12), 1382–1396. https://doi.apa.org/fulltext/2020-77456-001.html
17. Risk prediction of covid-19 related death and hospital admission in adults after covid-19 vaccination: national prospective cohort study*BMJ*2021;374:n2244: Authors: Hippisley-CoxJ,CouplandC A,MehtaN,KeoghR H,Diaz-OrdazK,KhuntiKet al. https://doi.org/10.1136/bmj.n2244
18. Financial Fragility in the COVID-19 Crisis: The Case of Investment Funds in Corporate Bond Markets; Authors: Falato, Antonio and Goldstein, Itay and Hortaçsu, Ali, doi: 10.3386/w27559,URL: http://www.nber.org/papers/w27559
19. Effects of COVID-19 Shutdowns on Domestic Violence in US Cities, Authors: Miller, Amalia R and Segal, Carmit and Spencer, Melissa K, doi:10.3386/w29429, URL: http://www.nber.org/papers/w29429
20. Financial Incentives and Other Nudges Do Not Increase COVID-19 Vaccinations among the Vaccine Hesitant, Authors: Chang, Tom and Jacobson, Mireille and Shah, Manisha and Pramanik, Rajiv and Shah, Samir B, doi: 10.3386/w29403, URL: http://www.nber.org/papers/w29403
21. An Efficient K-means Clustering Algorithm for AnalysingCOVID-19, Authors: Md. Zubair, MD.AsifIqbal, Avijeet Shil,Enamul Haque,Mohammed MoshiulHoque,Iqbal H. Sarker, https://doi.org/10.48550/arXiv.2101.03140