Clustering of Nationwide COVID-19 Data

Shaaf Afzal
Electrical Engineering
Washington University in St.Louis
St. Louis, Missouri
a.shaaf@wustl.edu

Washington University in St. Louis
St. Louis, Missouri
f.raynah@wustl.edu

Computer Engineering
Washington University in St. Louis
St. Louis, Missouri
l.trang@wustl.edu

Raynah Fandozzi
Systems Science & Engineering

Tri-Trang Le

Effective data analysis is crucial extracting meaningful insights and drawing reliable conclusions from complex data sets which can enable informed decision making in solving real-world problems. In this project, we developed a clustering algorithm to sort ambiguous counties' COVID data into the correct census division. In America, there are 9 divisions each exhibiting distinct COVID patterns, due to geographic and legislative factors. Initially, we sorted based on their states, as the variations in COVID responses, such as quarantine measures, vaccination and business closures, which varied across states. By analyzing both weekly new cases and the rate of new cases, we applied k-means clustering and compared their results to determine division with high accuracy. Our findings show that state-level legislation strongly impacts COVID trends, which aided in division identification. Further analysis incorporating vaccination rates, COVID testing data and deaths could further improve clustering accuracy and provide deeper insights to these COVID trends.

I. Introduction

In this case study, we are trying to understand the approach and reasoning behind clustering and classifying COVID-19 data for counties. Our primary goal is to develop a k-means clustering system that assigns counties to their correct divisions based on COVID-19 data trends while minimizing bias in the algorithm. A key requirement of the system is its ability to correctly classify new, unseen county data into the appropriate division. Ultimately, the aim is to create the best possible algorithm that maximizes accuracy and efficiency.

I. Methods

A. *Plan of Attack*

Our plan of attack is to use two sets of data to come up with a successful clustering mechanism. The first data set is the original COVID data [2] that we will smooth and normalize, but the second is the rate of COVID cases. This is calculated by finding the difference between two consecutive weeks. For example, the first column in that matrix would represent the rate of change from the first week to the second week of covid cases. We did this because the rate of covid cases is related to the geographic location. Then, our plan was to have an 80%/20% split for training and testing data respectively for both data sets. The goal of the training and testing data is to train our clustering algorithm, and then test its accuracy with an unbiased, random sample. The goal of the testing data is to determine if the clustering is unbiased, and able to successfully cluster data that it was not created with. To create these clusters, we used the k-means function in MATLAB.

B. *Noise Reduction*

The COVID data we have has a lot of noise. Since we are only looking at trends and signatures in the data, we are trying to reduce this noise as much as possible so that the sorting algorithm is able to more accurately sort the data. Initially, when looking at a graph of the data, we realized that the first major spikes and trends in the data show up after the 25th week and end at about the 135th week. So, we removed all that data. But this did not have a major impact on the accuracy rate. So, we trimmed more and trimmed the data in between the large peaks. For example, we trimmed the data from about week 60 to about week 70. But, when doing this, we ran into 2 issues. First, was the fact that our accuracy rate was not increasing significantly, and second, that trimming data was never a good thing. Although there are still trends in trimmed data, losing data is not always the answer. So, to successfully rescue the noise, we used the movmean function that we used in homework 2 to reduce the noise and used a k value similar to the one we used in the lab since it accomplishes a similar goal.

C. *Data Normalization*

COVID-19 cases are stored as raw counts (in hundreds of thousands of people), which vary across counties. For instance, a county in NY may have significantly more covid cases than a county in Vermont. As a result, we have to ensure that clustering is based on relative trends rather than absolute numbers, so we have to normalize the data. We decided to use a z-score normalization for the data [3]. We did this by finding the mean and standard deviation for each week for all the counties, then for each county's datapoints over the 156-week span, we subtracted the mean from each data point and divided each by the standard deviation of that specific week. This standardization process ensures that all counties are treated uniformly, regardless of their overall case volume. To interpret the data, a positive value (z-score) means that county has higher than average cases, but a negative z-score means that county has lower than average cases

D. Data Sorting

Before clustering, we realized that to effectively sort the data into meaningful clusters, the data must be organized in some way when first entering the k-means function. Initially, we sorted the counties by division, then took 20% of each division for testing, and the other 80% of each division for training. The issue with sorting the data like this was that the k-means centroids would be inherently biased. Meaning that the goal of this case study is to create unbiased centroids that will successfully sort any set of COVID data into division. But, if we were to have gone with the original plan of using the divisions to organize the clustering, it would have made the centroids biased to this set of covid data. We realized that we needed to the clustering to be based on other trends that the COVID data shared, which led to a correlation between these trends and division. So, we decided to initially sort the COVID data into states. We sorted the data into states and took about 20% of the values from each state to be put into the testing group and the rest of the values into the training group. (See III A and Fig. 1)

E. Kmeans

In our analysis of the COVID data, we used the k-means function to compute the centroids. In the k-means function, we initially used cosine distance between the counties to cluster them. We believed that the cosine distance function was useful in this context because it focuses on the orientation of the data points rather than magnitude. However, we realized that this difference in magnitude of population is a crucial factor in clustering. Therefore, we used Euclidean distance instead.

To increase the validity of our clustering, we used the replicates function to run the k-means 15 times with different initial centroids to determine the lowest sum of the differences. We did this for both the covid data and the rate of covid data sets. To determine the number of centroids, we determined the case in which our J value would be at a maximum.

$$J = N_{correct} - 0.5N_{centroids}$$

Where $N_{correct}$ refers to the number of correctly sorted counties and $N_{centroids}$ refers to the number of centroids.

One key consideration throughout the clustering process was the relationship between the number of clusters and the divisions of the counties, and how the different divisions could be represented. A common issue we ran into is divisions that are close in proximity, such as New England (division 1) and Middle Atlantic (division 2). The proximity and shared attributed of these regions led to clusters that blurred the boundary between divisions. Incorporating additional data which would further refine the analysis and be able to distinguish between regions that are geographically close but differ in other ways.

F. Determining Division from Clusters

To determine which cluster in each data set the random county will be associated with, we will use the nearest neighbor method using the cosine distance. This is unlike the clustering method which used Euclidean distance, but removing magnitude difference is crucial in this nearest neighbor method and correctly assigning counties to the correct cluster. For each cluster, we will have a top 3 divisions in it. More specifically, when we are trying to determine the division associated with each cluster, it is not a clear-cut answer. Meaning there can be some overlap between which divisions that cluster represents. To negate this, our plan is to have a top 3 (if there are even that many) divisions, in descending order, for each cluster in both the regular and rate covid data and then compare each county in them. Descending order means that the first division shows up the most frequent and so on. For example, we will have all my clusters trained and will be looking at a county in cluster 13 for the normal covid data, and cluster 5 for the covid rate data. Cluster 13 in the normal covid data is associated with divisions (2 5 8) while cluster 5 in the covid rate data is associated with only division 5. These two will compare themselves and determine that this county should most likely be in division 5.

II. Results and Discussion

A. Data Sorting: In Context

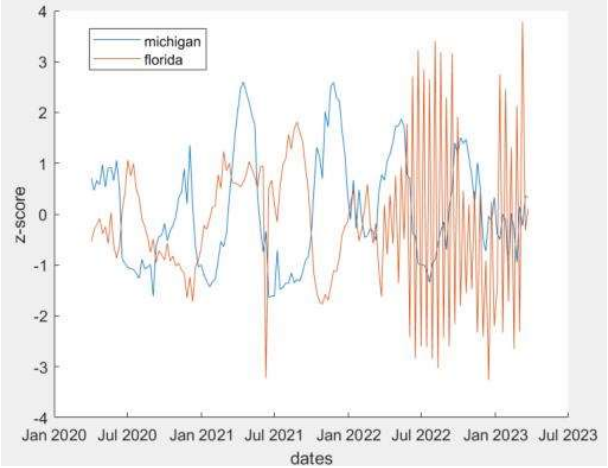

Fig. 1: COVID Cases of Michigan and Florida

For a little in-context explanation, we discuss the reasoning behind organizing the data by states to better train the kmeans. The question is why does the specific state influence the COVID data? That is because each state determined the response to COVID individually. For example, in the image above, we use Michigan and Florida as examples as 2 extremes between COVID responses. On one hand, Michigan had a fast and extreme response to COVID, closing all dine-in restaurants and making all non-mandatory workers work from home. The mandatory workers were regularly required to take COVID tests and get vaccinated. Also, the mask mandates were heavily in place. Unlike Florida, who essentially continued life as if nothing had happened, not really closing any restaurants or enforcing a mask mandate. In turn with this legislation, these two states COVID cases changed drastically. This graph shows the largest counties in Michigan and Florida. Looking at this normalized graph, you can see that initially, before any major legislation was enacted, the cases in the 2 counties acted similarly, spiking between a greater than average number of cases and a less than average number of cases. But, as time went on at about July of 2022, there is a clear difference in magnitude between Florida's and Michigan's peaks. Although Florida has a fluctuation in cases being both above and below the average, the peaks above average are much higher and more frequent than in Michigan. Towards the right of the graph, you can see this the most where Michigan's cases flatten out, but Florida's cases continue to be rampant. We assume that this is due to the state's legislation on COVID and may have something to do specifically with COVID vaccine mandates. This is one example of a constraint of our data. In this graph, we can see that initially, they have similar deviations from the average, but I believe this could be due to a lack of COVID testing in Florida. If we were to have access to the amount of COVID tests used/purchased by the state/county and we can view this graph in relation to the percentage of positive cases from overall tests, then the data could be different. If there were to be more tests purchased by Michigan, and more used by Michigan, which is my assumption then this would account of the lack of data initially in Florida. If we were to have this data, we would assume that the initial cases for Florida would be larger in their positive z-score.

B. COVID Rate: In Context

We calculated the COVID rate to be able to have a second set of data to analyze and compare with. The rate of COVID data is significant because the rate of COVID spreading is related to the geographic location. For example, if we look at the New England division compared to the Mountain division, the population density is a key difference. The population density in the New England division is greater than that of the Mountain division, meaning COVID will spread faster since there are more interactions between people. This will manifest in the COVID rate data and can serve as yet another way to be able to differentiate the divisions from each other. In our code, using both the clustering results from the regular COVID data and the clustering results from the COVID rate data will lead to a more accurate analysis.

C. Potential Improvements:

Additional data could help make our algorithm even more effective. For instance, given the deaths caused by each state, the relationship between cases and deaths could reveal the severity of COVID in specific counties. This ratio between deaths and counties would allow us to train our k-means more effectively given that we have more relevant information. As a result, our k-means could sort our testing data based on the severity of COVID in that county, sorting those with similar ratios within the same clusters and hopefully sorting the counties more effectively in accurate divisions.

Likewise, incorporating vaccination data for each county could significantly enhance the accuracy of our k-means algorithm when assigning counties to divisions. Since vaccination rates vary widely between states, this data would serve as a crucial feature. For instance, counties with higher vaccination rates are likely to experience fewer COVID cases due to increased immunity within the population. By calculating the percentage of vaccinated individuals in each county (vaccinated individuals divided by total population), we can track immunity levels. This additional feature would allow the algorithm to better differentiate between counties, creating more meaningful clusters and improving the accuracy of county-to-division assignments.

Another factor to make this approximation more successful is the average temperature for each country across different seasons. For example, the relationship between temperature changes and COVID-19 case rates may reveal trends where cases decrease in colder months due to reduced social interaction. By integrating temperature data into the k-means model, we could train it to recognize regions with seasonal fluctuations, improving the clustering process. The algorithm could then account for these environmental factors, grouping counties with similar temperature-related case trends, leading to more accurate sorting of counties into divisions.

Additional data on COVID-19 testing could further improve the effectiveness of our algorithm. By examining the number of tests conducted in each county, we can assess the testing rate relative to the population. With this information we can better measure the level of virus transmission within each county. For instance, a county with a high number of tests and a low number of cases might indicate effective surveillance and control measures, whereas a low testing rate with high case numbers could signify underreporting or a lack of access to testing resources. We can then again keep track of these ratios and incorporate them into our k-means algorithm. Incorporating this testing data into our k-means algorithm would allow us to create more detailed clusters, grouping counties based on their testing rates and case prevalence. This, in turn, would lead to a more accurate representation of how well each county is managing the pandemic and sort similar counties into the same division.

As mentioned earlier, incorporating data on government responses to COVID-19 could also improve the effectiveness of our clustering algorithm. By analyzing the timing of interventions, such as lockdowns, mask mandates, and vaccination campaigns, we can better understand their impact on COVID-19 case trends across different counties. For instance, counties that implemented early measures may exhibit different case rates compared to those with delayed or minimal responses. By including data that measures the timing of government responses, we can train our k-means algorithm to differentiate counties based on their response effectiveness. This additional layer of data would improve the clustering accuracy.

Another additional piece of data to consider is that of different variants of COVID-19 and the percentage of cases caused by these specific variants as well as details about its spread. This is an important piece of information to consider because it helps us understand the prevalence of this variant and how it affects the data. For instance, in week 100, there was a spike in the COVID-19 cases. This spike was due to the Omnicron variant, which became prevalent near this time. The

Omnicron variant is known for its high transmissibility and ability to evade immune responses, this insight can help us make our algorithm more effective. We know counties with higher populations will have more COVID-19 case data caused by the Omnicron variant and therefore we know population of counties should hold greater weight in our k-means clustering approach. Additionally, considering that humans' immune systems face more challenges in colder climates, we should also prioritize temperature in our model. By adjusting the weighting assigned to population and environmental factors, we can create a more accurate clustering algorithm.

Altogether, incorporating all these additional pieces of data to train our k-means would allow for more nuanced correlations between counties and clusters in the training process, ultimately leading to much more accurate results.

III. Conclusion

In summary, this case study shows the effectiveness of using k-means clustering to accurately group counties into correct census divisions based on COVID trends. This highlights the effectiveness of machine learning in classifying ambiguous data with a minimal bias. A key finding from this project is the significant influence of state-level legislation on COVID patterns. The clustering results reflected these variations, confirming that legislative actions and in turn, geographic location plays a large role in shaping the COVID spread. The correlation between COVID trends and the geographic location of counties provides valuable insights, especially considering that we only analyzed COVID case data.

Even though we were able to accurately cluster this COVID data based on division, the data is limited and potentially biased due to factors like differences in state and county-level responses, including potential disparities in testing protocols. These disparities can lead to inaccurate COVID data, as counties with limited testing may underreport data, while ones with more widespread testing could overreport case counts. To draw deeper conclusions, incorporating more data such as legislative measures, vaccination rates, death rates and environmental factors like temperature would lead to more accurate clustering and a more comprehensive analysis of this COVID data.

References

[1] United States Censys Bureau, County Population Totals and Components of Change: 2020-2023, 2023.

[2] The New York Times, Coronavirus (Covid-19) Data in the Unites States, 2023.

[3] Z. Bobbitt, "Z-score normalization: definition and examples," Statology, 12 August 2021.