COVID-19 Risk Assessment: Preparation and

Response

Mayuresh Hooli   
Technology & Solutions  
Ericsson Inc.Plano, USA  
[mayur.hooli@ericsson.com](mailto:mayur.hooli@ericsson.com)

Mason Jerome  
ECEN  
Texas A&M UniversityCollege Station, USA  
mmjerome15@gmail.com

*Abstract*— COVID-19 has been declared a global pandemic. Though the mortality rate is slightly lower than 5%, the rate of spread is unprecedented since the Spanish Flu in the early 20th century. The registered number of cases in USA was 174,684 as of April 1, but this number is set to see a sharp increase as the rate of cases has jumped significantly. With a limited number of test kits available, we need to identify the counties that are hotspots for the spread of the virus for proper allocation. Once the hotspots are identified, we need to identify the rate of spread to neighboring counties via our Machine Learning and Recommendation System techniques. If these critical parameters can be identified, then there will be some measures in place that will be crucial in fighting this virus. This approach helps to optimize the use of the kits and addresses the most important issues we have right now through a disaster management approach. As time passes, we can evaluate the results of our approach with the actual value of cases in each county in Texas. We extract some meaningful features out of public census datasets and try to identify the hotspots. We will use data analytics and statistical analysis to study the effect of each indicator on spreading the virus and predicting its spread into hotspots. This idea can help local and state governments to logistically optimize their resource allocation.

Keywords—COVID-19, Ranking (Statistics), Clustering, Information Retrieval.

# Introduction

Coronavirus disease 2019 (COVID-19) or Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a disease closely related to SARS-CoV which was responsible for the SARS epidemic of 2003 [1]. Coronavirus belongs to a group of viruses that infects various animals like dogs and bats. The name coronavirus originates from the bulb-tipped spikes that project from the surface of the virus and give the appearance of a corona surrounding it. The coronavirus infection tends to play out in two ways: either as an injection in the lungs that would cause common cold or an infection in the guts that would cause diarrhea. This starts out in the lungs like a common cold, but then wreaks havoc on the immune system in some cases resulting in long-term lung damage amounting to death. While this bears similarities to the SARS-CoV, there are over 6000 mutations sprinkled around [2].

It is difficult to determine the origin of COVID-19, though it is improbable that it emerged from laboratory manipulation of SARS-CoV. Two scenarios can plausibly explain the origin of COVID-19: (i) natural selection in an animal host before zoonotic transfer; and (ii) natural selection in humans following zoonotic transfer.

## Natural selection in an animal host before zoomotic transfer

The early cases of COVID-19 were related to the Huanan market in Wuhan which postulates an animal source present at this location. As the SARS-CoV-like virus in bats was similar to the SARS-CoV-2 virus, it is likely that bats were the initial hosts of this virus. Although this is ~96% similar to SARS-CoV-2, recent studies have suggested that its spike diverges in RBD, which suggests it may not bind efficiently to human ACE2. Another potential source is the Malayan pangolin illegally imported in to the Guangdong province which contain coronaviruses similar to SARS-CoV-2 [3].

## Natural selection in humans following zoonotic transfer

This theory states that the progenitor of SARS-CoV-2 jumped into humans, acquiring the genomic features described above through adaptation during undetected human-to-human transmission. Once the genomic features were acquired, these adaptations would enable the pandemic to take off and produce a sufficiently large number of cases.

The COVID-19 Management is an intelligent system which identifies the critical areas (case study: Texas) where certain actions need to be taken. This system also ranks the counties based on the risk they pose to other counties for spreading the virus. This novel idea is based on implementing clustering alongside the PageRank algorithm. Clustering identifies 4 clusters based on the data set we prepared. PageRank algorithm uses the clustering prediction to identify which counties risk spreading the virus further. There is a shortage of medical personnel and testing kits, the categorization of counties helps in effective allocation of resources. The prediction of future affected counties will help in better preparedness to mitigate the virus. Ericsson is a pioneer in the field of communication. The communication technology developed by Ericsson would help understand the travel history of a county on an average to better prepare for the coming weeks.

# Proposed Method

The idea is more useful for state and local government in terms of optimizing resources in this difficult time, but it can be scaled for use on a federal level. The Disaster Management tool can potentially get funding from the government. The government could benefit from its analysis to better control the distribution of test kits and implement policies to reduce the spread of the virus.

## Novelty

Policymakers around the world are trying to figure out how to fight Coronavirus. The novel approach of using Clustering and Page Rank together can help them understand where the risk lies. Following the results given by this combination of algorithms would help them manage the fight against the pandemic.

*B. Clustering*

Clustering is a technique that divides data points into different groups based upon their features. Using this method provides clear groupings for the counties that lets us classify them into as many clustered segments as chosen. For the purpose of this dataset, we decided to delineate the counties into 4 different clusters with K-means clustering. This algorithm finds datapoints closest to each assigned cluster centroid and iterates, updating the cluster centroid locations until error is minimized to find the best cluster assignments. Upon arriving at the final clusters, the PageRank algorithm is used to determine the severity of COVID-19’s impact upon each of the counties from other counties.

*C. Page Rank Algorithm*

PageRank is an algorithm originally developed by Google to rank web pages. For the purposes of this project, we slightly modified the algorithm to identify the impact that counties have on each other in regard to the spread of COVID-19. This is done by taking the previously determined severity rank value of each county and transferring a portion of that value to neighboring counties based upon how many neighbors the county in question has. So, if a county has fewer neighboring counties, it will assign a larger portion of its value to add to those counties, and vice versa. Understanding inter-county spread of the COVID-19 virus can help inform local and state governments how to tackle the difficult problem of curbing the virus’ spread.

*D. Data*

Meaningful features are extracted from public census datasets to identify hotspots. We will use data analytics and statistical analysis to study the effect of each indicator on spreading of the virus and predicting the spread into neighboring hotspots.

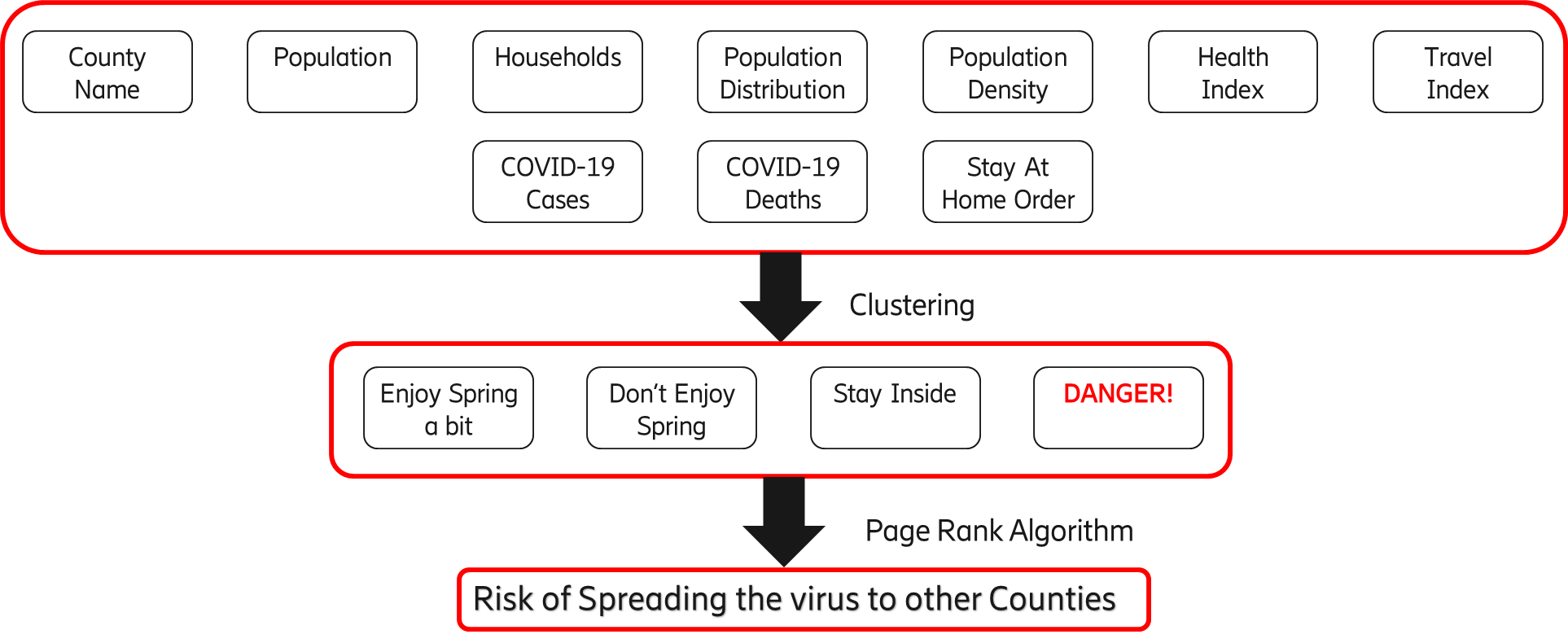


Fig. 1. The general view of our proposed methodology.

Here, we predict the danger of COVID-19 based within the counties. This can be scaled to a larger level for states and countries as well. With a more specified set of census data, we can better predict the spread of COVID-19 based on the counties. Multiple data sets are combined to get the final, aggregated data set required for this Disaster Management tool such as census, cleanliness, and COVID-19 data.

The census data consists of the Geo ID of the county, FIPS, county name, the number of households, and the total male and female population as well as the population based on the gender percentage demarcated by 5-year age groups. We combine the total population categories for the age ranges of 0-15, 15-65 and 65+ for the data set. We do not consider the gender ratio for the data set.

The next data set we include is the health index. The health index is derived from the healthiest communities’ data set on US News. This data set contains information about the population health, equity, education, economy, housing, food nutrition, environment, public safety, community vitality, and infrastructure of the county. We aggregate all this data to calculate the cleanliness index of the counties to use in the data set. As we do not need the rest of the data now, we delete those columns and only use the aggregated column.[4]

Next, we obtain the population density from the land area. This is used due to decreased social distancing and higher rates of contact between people in places where the population density is higher. The population density is aggregated for the county rather than considering the concentration in a few places that is seen in various counties.

The next parameter considered is the travel index. The travel index constitutes a key feature in determining inter-county spread of COVID-19. It was derived by calculating the correlation coefficients between the likelihood of owning a passport and median income (m), economic output (p), and degree of education (e). Once the correlation coefficients are known, a simple formula we came up with can derive the travel index values for each county.[5]

For a singular example of a county, given normalized outputs of:

m = 0.81

p = 0.7

e = 0.7

*Tind=Nm×0.81+ Np×0.7+ Ne\*0.73*

This travel index is then generalized to indicate the frequency of travel, indicating a greater degree of contact and spread of the virus. I then follow that the counties with higher travel indices have a greater chance of spreading the virus to neighboring counties. Putting focus on these high travel index counties could yield proportionally higher results in attempts to curb the spread of the virus in comparison to counties with a low travel index, making it an important feature to include in the clustering algorithm.

The next parameter we consider is the COVID-19 cases and deaths in each county. We also use the COVID-19 cases and deaths by population and household. Since these numbers are very small, we normalize them to get something we can use. We also use the COVID-19 stay at home labels. Since many of the Texas counties are rural counties with a small population density, people tend not to interact much with each other. We give the value 0 for counties that haven’t implemented stay-at-home orders and 1 for counties that have implemented the stay-at-home orders.

*F. Experimental Results*

Using the parameters detailed above, we implement the K-means clustering algorithm. We group the counties into 4 different clusters. The following diagrams give the results of the clusters in table I.

*G. Spread Identification*

Now we have identified the counties which are hotspots for the COVID-19. However, there may be more counties that are at risk of becoming COVID-19 hotspots in the future. We base this on the proximity of the counties to each other. The counties that have a high concentration of COVID-19 cases may have an adverse impact on other counties and raise their number of COVID-19 cases. To identify the threat posed by other counties to the county in question, we use the PageRank algorithm. To use the PageRank algorithm, we use the matrix of neighbors. But as we see above, not all the neighbors are equal. Thus, we replace the value 1 with an addition of cases per unit area and deaths per unit area and then take a square root of them. Then we normalize this value to get the values in the matrix. We apply a PageRank algorithm to this matrix. Based on this matrix, we observe that the Dallas county has the highest risk of being infected even further while Wheeler County is at the lowest risk.

|  |  |  |
| --- | --- | --- |
| Texas Counties | COVID-19 cases reported | Labels |
| Aransas | 2 | Enjoy Spring a bit |
| Archer | 0 | Enjoy Spring a bit |
| Austin | 2 | Enjoy Spring a bit |
| Bandera | 0 | Enjoy Spring a bit |
| Baylor | 0 | Don’t Enjoy Spring |
| Elpaso | 40 | Don’t Enjoy Spring |
| Hidalgo | 28 | Don’t Enjoy Spring |
| Wichita | 28 | Don’t Enjoy Spring |
| Nueces | 22 | Don’t Enjoy Spring |
| Cameron | 20 | Stay Inside |
| Travis | 200 | Stay Inside |
| Denton | 165 | Stay Inside |
| Collin | 134 | Stay Inside |
| Fort Bend | 119 | Stay Inside |
| Galveston | 70 | Stay Inside |
| Bexar | 157 | Danger |
| Dallas | 488 | Danger |
| Harris | 526 | Danger |
| Tarrant | 139 | Danger |

TABLE I EXPERIMENTAL RESULTS

III. CONCLUSION & FUTURE WORK

This method of combining PageRank algorithm with clustering is an innovative technique to fight COVID-19. Clustering helps us identify the current danger zones while PageRank provides us information on the future danger zones. The initial observations of this experiment proved that this technique is accurate and could be easily expanded further.

For future work, we plan to collect the cell phone data to better understand the travel history and aggregate on counties to extend this algorithm to the whole. We will use the Geographic Information System to identify the zones on the map for better visualization. We will combine model with time series analysis. We will create a similar model for pandemic response and disaster response to help manage similar situations in the future.

REFERENCES

[1] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7438984/>

[2] Jiang, Dong, et al. "Mapping the transmission risk of Zika virus using machine learning models." *Acta tropica* 185 (2018): 391-399.

[3] Andersen, Kristian G., et al. "The proximal origin of SARS-CoV-2." Nature medicine 26.4 (2020): 450-452.

[4] https://www.usnews.com/news/healthiest-communities/texas

[5] <https://www.theatlantic.com/national/archive/2011/03/americas-great-passport-divide/72399/>