Olivia: Data Sources and References

I. COVID-19 Daily Information per County

Our first step towards the mentioned objective is to gather the daily COVID-19 outbreak data. This data should include the number of cases that are confirmed to be caused by the novel coronavirus and its associated death toll. We are using the publicly accessible dataset API in [1], [2] to fetch the relevant data records. The table of data obtained using this API contains the numerical information along with dates corresponding to each record, and each document includes the number of confirmed cases and the number of deaths that occurred due to COVID-19 on that date. It also includes the number of recoveries from COVID-19 in the same format. This dataset's significance is that it provides us with a detailed and high-resolution temporal trajectory of the COVID-19 outbreak in different urban regions across the United States. Using the dates, one can constitute a set of time-series for every county and monitor the outbreak along with the other metadata to make relevant inferences.

II. US Census Demographic Data

The US Census Demographic Data gathered by the US Census Bureau [3] plays a critical role in our analysis by providing us with necessary information on each region's population. Additionally, this information includes specific features such as the types of work people in that region mainly take part in, their income levels, and other invaluable demographical and social information.

1) US County-level Mortality: The fluctuations in the mortality rate of a region is also a potential critical feature in pandemic analytics. The US county-level mortality dataset was incorporated into our collection to add the high-resolution mortality rate time-series throughout the years [4], [5]. The age-standardized mortality rates provide us with information on variables, the values of which can be considered as the effects of specific causes. It is crucial since some of these causes might have contributed to the faster spread of COVID-19 in different regions [6].

III. US County-Level Diversity Index

Another dataset that offers a race-based breakdown of the county populations is available at [7] with the diversity index values corresponding to the notion of ecological entropy. For a particular region, if K races comprise its population, the value of diversity index can be computed using the following formula:

$$d_i = 1 - \sum_{i=1}^{K} (\frac{n_i}{N})^2$$

In the above formula, N is the total population and n_i is the number of people from race i. This formula represents the probability p, which means that if we randomly pick two persons from this cohort, they are of different races with probability p. In addition to that, we have the percentages of different races in the regional population as well.

IV. US Droughts by County

Another source of valuable information regarding the land area and water resources per county is the data gathered by the US drought monitor [8], [9]. This data is incorporated into our collection as well.

V. Election

Based on the 2016 US Presidential Election, a breakdown of county populations' tendencies to vote for the main political parties is available [10]. These records are added to our collection as the democratic-republican breakdown of regional voters can reflect socio-economic and demographical features that form the underlying reasons for the regional voting tendencies.

VI. ICU Beds

Since COVID-19 imposes significant problems in terms of the extensive use of ICU beds and medical resources such as mechanical ventilators, having access to the number of ICU beds in each county is helpful. This information offers a glance at the medical care capacity of each region and its potential to provide care for the patients in ICUs [11]. It could be argued that having knowledge of the ICU-related capacity of regional healthcare providers can, to some extent, represent the amount of their COVID-19 related resources, such as ventilators and other needed resources.

VII. US Household Income Statistics

The aggregate dataset on central statistical values on the US household income per county (including average, median, and standard deviation) is used to provide information on the financial well-being of the affected regions' occupants [12].

VIII. COVID-19 Hospitalizations and Influenza Activity Level

Aside from the socio-economical and demographical features of a region, the number of active and potential COVID-19 cases is a critical factor. This information can be leveraged to provide a possible threat level for the region. These records are made available by CDC for specific areas and are incorporated into our collection as well [13], [14].

IX. Google Mobility Reports

The COVID-19 virus is highly contagious. Therefore, the self-quarantine and social distancing measures are principal effective methodologies in bolstering the prevention efforts. Our collection includes Google's mobility reports obtained from [15]. These records elaborate on the mobility levels across US regions, which are broken down into the following categories of mobility:

1) Retail and Recreation
2) Grocery and Pharmacy
3) Parks
4) Transit Stations
5) Workplaces
6) Residential

In addition, we have computed a compliance measure that has to do with the overall compliance with the shelter at home criteria:

$$\text{compliance} = -1 - \frac{(1/6)\sum_{i=1}^{6} m_i - 100}{100.0}$$

In the above formula, m_i is the mobility report for the ith mobility category. This value is computed through time to provide an overall measure of mobility through time. The compliance measures of $+1$ and -1 mean $+100\%$ and -100% changes from the baseline mobility behavior, respectively.

X. Food Businesses

Restaurants and food businesses are affected severely by the economic impacts of this outbreak. At the same time, they have not ceased to provide services that are essential and required by many. To reach a proper perspective of the food business in each region, we have prepared another dataset based on records in [16] to provide statistics on regional restaurant revenue and employment. Analysis of restaurants' status is important in the sense that they are mostly public places that host large gatherings, and in the time of a pandemic, their role is critical.

XI. Physical Activity and Life Expectancy

Various features have been selected from the dataset in [17] to reflect on the obesity and physical activity representation for different US regions. These features include the last prevalence survey and the changes in patterns. Also, Life Expectancy related features are valuable information for representing each region. They are included as well in our analyses.

XII. Diabetes

Different features to represent a region according to the diabetes-related characteristics were selected from the data in [17]. These include age-standardized features and clusters that have to do with diabetes-related diagnoses.

XIII. Drinking Habits

Information on regional drinking habits from 2005-2012 has also been used in this work [17]. This information includes the proportions of different categories of drinkers clustered by sex and age. The categories are as follows:

- "Any": a minimum of one drink of any alcoholic beverage per 30 days
- "Heavy": a minimum average of one drink per day for women and two drinks for men per 30 days
- "Binge": a minimum of four drinks for women and five drinks for men on a single occasion at least once per 30 days

TABLE I
OVERVIEW OF THE FEATURES

Category	Description
Food Businesses (static)	Food and Beverage Locations Restaurant Employments Sale and Economy
Gender (static)	Percentage of Male and Female
Race (static)	Ratio of different races
Election (static)	Ratio of Democratic, Republican, and other voters
Income (static)	Wage Statistics Poverty Information
Commute (static)	Statistics of Methods of Commute to Work and Their Ratio
Hospitals and Mortality (static)	Information on ICU Capacity and Statistics on Region's Mortality
Obesity and Physical Activity (static)	Information on the Statistics of Obesity and Physical Activity and the Changes in Patterns
Life Expectancy (static)	Regional Life Expectancy Values in Years
Drinking (static)	Alcohol Consumption Patterns and Changes
Diabetes (static)	Patterns of Different Types of Diabetes Diagnoses and Changes in Them
Land and Water (static)	Information on Land and Water Resources of Regions
Employment (static)	Ratio of Different Job Types and Other Statistics
CDC Hospitalizations and Surveys (dynamic)	Number of Hospitalizations due to COVID-19 and Influenza Activity Surveys
Google Mobility Reports (dynamic)	Breakdown of Regional Mobility in Different Categories Based on Which Our Compliance Score Is Computed

REFERENCES

[1] "Covid-19/coronavirus real time updates with credible sources in us and canada," https://coronavirus.1point3acres.com/en, archived at https://archive.is/J3Vmg, accessed: 2020-06-05.

[2] T. Yang, K. Shen, S. He, E. Li, P. Sun, L. Zuo, J. Hu, Y. Mo, W. Zhang, P. Chen *et al.*, "Covidnet: To bring the data transparency in era of covid-19," *arXiv preprint arXiv:2005.10948*, 2020.

[3] "Us census demographical data," https://www.kaggle.com/muonneutrino/us-census-demographic-data, archived at https://archive.is/ZY12v, accessed: 2020-06-05.

[4] "Us mortality rates by county," http://ghdx.healthdata.org/record/ihme-data/united-states-mortality-rates-county-1980-2014, archived at https://archive.is/juSbk, accessed: 2020-06-05.

[5] "Us county-level mortality," https://www.kaggle.com/IHME/us-countylevel-mortality, archived at https://archive.is/xEVs3, accessed: 2020-06-05.

[6] L. Dwyer-Lindgren, A. Bertozzi-Villa, R. W. Stubbs, C. Morozoff, M. J. Kutz, C. Huynh, R. M. Barber, K. A. Shackelford, J. P. Mackenbach, F. J. van Lenthe *et al.*, "Us county-level trends in mortality rates for major causes of death, 1980-2014," *Jama*, vol. 316, no. 22, pp. 2385–2401, 2016.

[7] "Diversity index of us counties," https://www.kaggle.com/mikejohnsonjr/us-counties-diversity-index, archived at https://archive.is/uX9iX, accessed: 2020-06-05.

[8] "Us drought monitor," https://droughtmonitor.unl.edu/, archived at https://archive.is/P76Bb, accessed: 2020-06-05.

[9] "United states droughts by county," https://www.kaggle.com/us-drought-monitor/united-states-droughts-by-county, archived at https://archive.is/JgGoj, accessed: 2020-06-05.

[10] "County presidential election returns2000-2016, 2018," https://doi.org/10.7910/DVN/VOQCHQ, archived at https://archive.is/cLVL5, accessed: 2020-06-05.

[11] "Icu beds by county in the us," https://www.kaggle.com/jaimeblasco/icu-beds-by-county-in-the-uss, archived at https://archive.is/QgAoO, accessed: 2020-06-05.

[12] "Us household income statistics," https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations, archived at https://archive.is/iJaLT, accessed: 2020-06-05.

[13] CDC, "A weekly summary of us covid-19 hospitalization data," https://gis.cdc.gov/grasp/COVIDNet/COVID19_1.html, archived at https://archive.is/qs0IJ, accessed: 2020-06-05.

[14] ——, "Laboraty-confirmed covid-19 associated hospitalizations," https://gis.cdc.gov/grasp/covidnet/COVID19_3.html, archived at https://archive.is/Mw9d1, accessed: 2020-06-05.

[15] B. D. Killeen, J. Y. Wu, K. Shah, A. Zapaishchykova, P. Nikutta, A. Tamhane, S. Chakraborty, J. Wei, T. Gao, M. Thies *et al.*, "A county-level dataset for informing the united states' response to covid-19," *arXiv preprint arXiv:2004.00756*, 2020.

[16] N. R. Association, "State statistics," http://web.archive.org/web/*/https://www.restaurant.org/research/state, accessed: 2020-07-15.

[17] "Us data for download," http://web.archive.org/web/20200121125528/http://www.healthdata.org/us-health/data-download, accessed: 2020-04-02.