Data Challenge Methods

I.Data Acquisition

In pursuing our research question, we first created a list of priority data to obtain, which included demographics, unemployment data, health care coverage, hospital location and capacities, voting outcomes by county across the US, flight data, comorbidities, and air quality data. County level COVID - 19 cases and deaths, as well as county level shapefiles were provided by the course instructor. All other data was found via google search using keywords such as: “Voting”, “Unemployment”, “Health care”, “Transportation”, “State Funding”, “COVID-19 Impacts”, and “Flights”. Data that were reported at the county level were unified using FIPS codes. If the data did not have FIPS codes upon download, they were found and added to files. We created a data registry to keep track of all acquired data along with key characteristics of the data seen below.

II. Data Pre-processing

To make the data analysis ready, data were processed in R Studio (R Core Team, 2020) by choosing key variables from each dataset, converting it to “long” format, and merged all datasets with a cross-walk of state and county names and FIPS codes to ensure spatial consistency among variables for incorporation into models and visualizations. All variables were labeled as either predictor or outcome variables and all possible combinations of the two variable types were run through simple linear regression models to plot pairwise scatterplots. From this, correlations were explored to determine which variables would be of greatest significance to our logistic regression model and spatial autocorrelation analysis. Predictor variables used for this pre-analysis were: Number of hospitals and beds, median AQI (air quality), community resilience, non-pharmaceutical response measures implemented by state governments, demographic data from the Census Bureau’s American Community Survey (ACS) about race, income, age, use of public transportation, and portion of the population with health insurance, comorbidity, and percent of votes earned by the democratic candidate by county in 2016. Outcome variables used in the simple linear regression were: change in unemployment at the state and county level, COVID- 19 reported cases per capita, and COVID-19 reported deaths per capita.

III. Data analysis/model building and implementation

We aggregated the total number of cases and deaths attributed to red states vs. blue states (with redness/blueness defined by 2016 state election results or by the party affiliation of the governor). Then, we constructed a bar plot to illustrate how the United States cases and deaths compared to other countries after accounting for the contributions of blue states. Box plots, scatter plots, and linear regressions were generated to examine differences between red and blue states in the 2016 election (predictors) and COVID-19 cases, death counts per capita, and unemployment rate (responses) at county and state level. Redness/blueness was assessed by the percentage of total votes for democratic candidates (Hillary Clinton) in a given county/state. All models exhibited weak relationships with the greatest adjusted R^2 of 0.1267. Thus, several additional predictors were examined in explaining the COVID19 outcomes using regression analysis and AIC model selection and random forest classification.

We next created correlation plots to assess the relationships between all dependent and independent variables at the county and state levels. At the county level, this included: air

quality, community resilience, and hospital data (number of beds, etc.). At the state level, relevant predictor variables included: hospitals per state, number of governmental measures in response to COVID-19, median income, and fractions of the population who identify as black, pacific islander, "other" races, under 18, over 65, and those who are insured. Several significant correlations were identified. Next, we constructed linear and/or generalized linear models to identify predictors of cases per capita, deaths per capita, and change in unemployment rate. The initial models considered all numeric and categorical variables as predictors. After running the first linear model, we removed all of the numeric predictors that were not significant at the 5% level. Variance inflation factors (VIFs) were determined to test for multicollinearity among predictors; variables with VIF values greater than 5 were removed from their associated model. AIC tests were applied to determine optimal predictors for each response variable, respectively. Individual predictor variables that were included in the optimal models that resulted from the AIC tests were then transformed by exploring typical transformations for each variable, i.e. log(variable), sqrt(variable), (variable^2), (1/variable), etc. Of these, the transformations that most significantly improved the fit of the model (if any) were kept. This led to an improvement of fit based on the plots (qqplot, histogram, residuals vs. fits, etc.) as well as the significance of each variable and adjusted R^2 value. However, these models did not account for spatial autocorrelation in the data and were thus not used as the final model.

Other preliminary models run were quasi-binomial logistic regressions, beta logistic regressions, and beta log-log regressions using the same variables that were included in the optimal linear model for each response variable. Each model was run thrice, once for cases per capita, deaths per capita, and change in unemployment rate, respectively. Of all approaches explored, the beta models outperformed all other models in nearly every case, but the R^2 was still extremely low.

The final model we used to answer our question was a random forest model. Random Forest Regression was used to identify features that contribute most to Covid Case Counts, Covid Death Counts, and Unemployment at county scale. We chose a random forest model because it is an ensemble model that is robust, and because there was a pretty large dimension to the data given. Initial analysis included Moran's I spatial autocorrelation analysis, and in all three cases (Covid Cases, Covid Deaths, and Unemployment) spatial autocorrelation was significant. The model was trained using a 70/30 split and validated with the full data set.

IV. Data Visualization

We used R to visualize our data and create figures that could be embedded on our website. Our code for these figures can be found in the Github repository in the master branch in the visualization folder. We created figures for both our input data (such as unemployment by state over time or COVID-19 rates at the county level) and our analysis output. We created non-interactive figures in R then exported to a *png* or *jpg* file. For interactive figures, we created them in R then exported them as html widgets which can also be embedded in our website.

We used github pages with Jekyll to summarize and present our findings and figures. We chose this medium due to its shareability and interactivity. Once we have finalized our analysis, the page can be made public and will succinctly display our analysis with links to our methods and github page. The page is located in the gh-pages branch of the main repository. The format for the page is a modified version of the Minimalist theme from Jekyll. These formatting modifications were made by Garret Miller (PhD Student in NC State's Center for Geospatial Analytics, https://gcmillar.github.io/) and our group.

Table 1 Data registry of all datasets collected

File name	Description	File type	Data source/link	Spatial Resolution	Temporal Resolution	Year(s)/ time-step	Used in Final Analysis?
COVID19_non-pharmaceutical-interventions_version2_utf8	Dataset of government interventions in response to COVID-19	csv	(US data was extracted from the total dataset) https://www.nature.com/articles/s41597-020-00609-9#Abs1	State	-	1/2020 - 5/2020	No
cb_2015_us_county_20m	Shapefile of all counties in U.S.A.	shapefile	Josh	County	2015	2015	Yes
statepres_1976-2016	State- level returns for elections to the U.S. presidency from 1976 to 2016.	csv	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/42MVDX	State	-	1976-2016	Yes
countypres_2000-2016	County-level returns for presidential elections from 2000 to 2016. (see line 22 of this doc)	csv	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VOQCHQ	County	-	2000-2016	Yes
covid_confirmed_usafacts	Table of confirmed COVID-19 cases by county in all 50 states	csv	Josh	County	Daily	1/22/2020 - 9/22/2020	Yes
covid_county_population_usafacts	Table of county population	csv	Josh	County	-	Unknown	Yes
covid_deaths_usafacts	Table of deaths from COVID-19	csv	Josh	County	Daily	1/22/2020 - 9/22/2020	Yes
US hospitals	Shapefile of locations and capacities of hospitals across US 50 states	shapefile	https://hifld-geoplatform.opendata.arcgis.com/datasets/6ac5e325468c4cb9b905f1728d6fbf0f_0	County	-	Current	Yes
COVID-19 sentiment	Polygon layer of public perception (sentiment) of COVID-19	shapefile	https://www.arcgis.com/home/item.html?id=feb6280d42de4e91b47cf37344a91eae	County	Weekly	12/2019 - 9/2020	No
American Community Survey (ACS)	American Community Survey	csv	R package - "acs"	County	Annually	2018	Yes
communityresilience_county	County level Community Resilience	csv	https://www.socialexplorer.com/data/CRS2020/documentation/	County	-	2018	Yes
communityresilience_state	State level Community Resilience	csv	https://www.socialexplorer.com/data/CRS2020/documentation/	State	-	2018	Yes
Legislative control spreadsheet	2020 State & Legislative Partisan composition	csv	https://www.ncsl.org/Portals/1/Documents/Elections/Legis_Control_2020_August%201.pdf?ver=2020-08-04-135320-640&timestamp=1596570819021	State	-	2020	Yes
air_quality_annual_aqi_by_county_2020	2020 Air quality data by county	csv	https://aqs.epa.gov/aqsweb/airdata/download_files.html	County	-	2020	Yes
air_quality_annual_aqi_by_county_2019	2019 Air quality data by county	csv	https://aqs.epa.gov/aqsweb/airdata/download_files.html	County	-	2019	Yes
analytic_data2020_0.csv	The County Health Rankings - snapshot of community health	csv	https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation	County	-	2020	No
comorbiditesbyage.csv	Conditions that contributed to deaths caused by COVDI 19	csv	https://data.cdc.gov/NCHS/Conditions-contributing-to-deaths-involving-corona/hk9y-quqm	State	-	2/2020 - 9/2020	Yes
Flight data	US ontime flight statistics from Dec 2019 - June 2020	csv	https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236	State	Daily	12/2019 - 6/2020	No
2016_US_County_Level_Presidential_Results	County vote totals in 2016	csv	https://github.com/tonmcg/US_County_Level_Election_Results_08-16/blob/master/2016_US_County_Level_Presidential_Results.csv	County	-	2016	Yes
county_unemploymentfebjuly2020	Unemployment statistics by county	csv	https://www.socialexplorer.com/data/US_unemployment_2020/metadata/?ds=ORG	County	-	2/2020 - 7/2020	Yes
WHO-COVID-19-global-data.csv	COVID19 cumulative cases and deaths by country	csv	https://covid19.who.int/	Global	Daily	12/2019 - 09/2020	Yes