Cover page

Team members:

1. Name: Longhao Jin (team lead), Netid: longhao2
2. Name: Yongchang Su, Netid: ysu17

Stat 542 final project report

Longhao Jin, Yongchang Su

1 Project Overview

This project tries to understand the reasons behind the massive outbreak of COVID-19 and aims at finding an effective way to prevent the transmission of the virus. Moreover, we hope this report can provide the public a guidance of how to protect themselves during this pandemic period. We conclude that the government should pay attention to the vulnerable people, especially for female and the older. Besides, each county should provide sufficient medical resources to the people who cannot afford it. Constructing the temporary hospitals is another way to help reduce the mortality. Last but not least important thing is that the public should follow the anti-contagion policy announced by the government. Although we do not observe the effect significantly important in our models, those policies will have a huge influence if we learn the well-organized lessons from China.

The rest of the report is organized as follows. Section 2 provides a literature review of the studies on COVID-19 topic. Data processing will be implemented in Section 3. Several unsupervised learning algorithm such as dimension reduction technique (PCA), spectral clustering, hierarchical clustering and K-means clustering will be performed in Section 4 to learn from the data-set. In section 5, we perform random forest, support vector machine and logistic regression in the classification problem. And similarly, linear regression with Lasso penalty, random forest and boosting method will be applied in the regression problem. We answer the collaborate question with concluding remarks in Section 6.

2 Literature Review

Altieri et al. (2020) tries to predict the expected number of death cases for the next 7-days by introducing different types of prediction models. There are 5 models proposed in the paper: (1) exponential model w.r.t time t in different counties; (2) exponential model w.r.t the log-transformation on the death records which gathered from all of the countries; (3) expanding the method in (2) to a much broader scenario, where the confirmed deaths and cases are considered among one county and the neighbors of it; (4) exponential model that take the additional demographic features into consideration; (5) changing the exponential setting to a linear model in scenario (1). Moreover, the paper also includes some ensemble approach, where the models mentioned above are combined with weights based on their

predictive performance when they are fitted separately. The result shows that when focusing on 7-day predictions, the combination of expanded shared predictors and linear predictors has the best performance. The idea is borrowed from the Schuller, Yu, Huang, and Edler (2002), where the weights are proportional to an exponential function of the loss function between the true value and the predicted value. Using the similar thought in (2) and (3), Elmousalami and Hassanien (2020) proposes a time series based prediction model to forecast the cases in a daily basis. The moving average, weighted moving average and single exponential smoothing model frameworks are performed.

3 Data processing

3.1 Missing Values

There are different methods to deal with the missing values, detailed discussion will be addressed in Appendix A.

1. Deleting the whole column.
2. Imputing the missing values by the mean of the column if the number of missing values is less than 5%.
3. Imputing the missing values by some policy-related assumption.
4. Imputing the missing values by using regression.

3.2 Combining the Columns

Define the following age groups:

Group	Baby	Child	Youth	Adult	Senior
Age	<5	5-14	15-24	25-64	>65

Then combining the age columns into new columns named by the above groups.

4 Unsupervised learning

4.1 Dimension Reduction

Before any clustering methods, we realize that the data containing demographical and health-related information can be severely correlated. Therefore, it will be reasonable to perform PCA to reduce the dimensionality of the data first.

As shown from the Figure 8 in Appendices, the first 5 principal components already contain 90% of the information, so we only use them for further analysis of the underlying clusters of counties.

We further examine the top 2 PCs, and find that first PC has high weights in magnitude on all population variables, so we call it population PC, with small value indicating larger

population. For second PC, it has high weights on most of the health-related variables, with small value indicating bad health condition.

4.2 Spectral Clustering

We perform spectral clustering on the data, and get the following results.

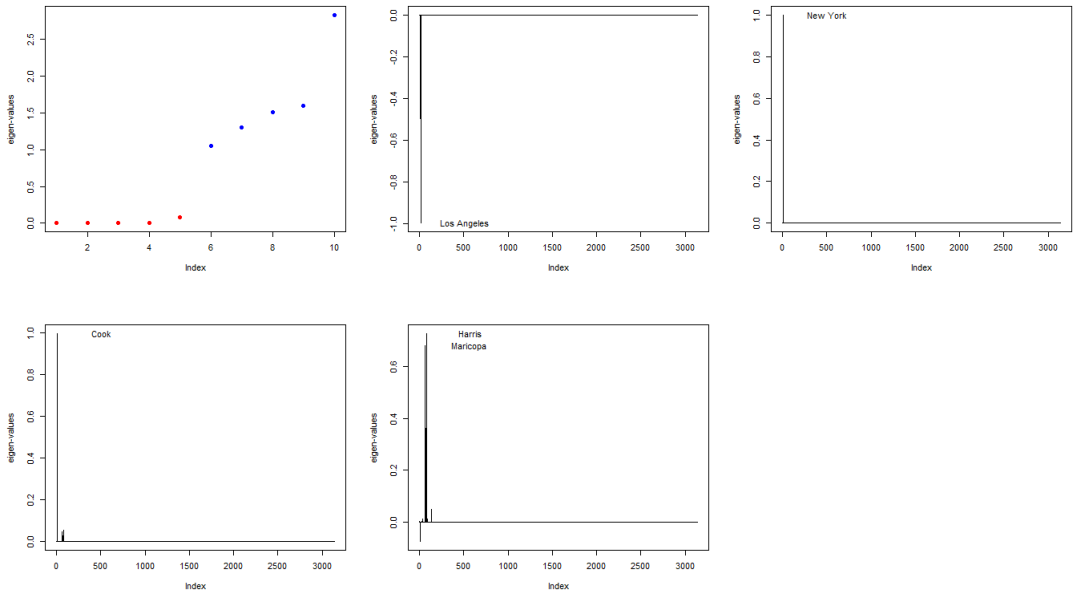


Figure 1: Spectral clustering on demographic and health

Based on Figure 1, it seems that the majority of US counties are similar to each other and falls into the same cluster. Only very few counties, like Los Angeles, New York, Maricopa, Harris and Cook are excluded from the main cluster.

Continuing on the analysis on top 2 PCs, from Figure 9, it seems that residents from medium-sized counties have better health condition than those from extremely large or small counties. This is a reasonable phenomenon since large counties may have trouble offering good medical care to all of its huge population and good health-related experts may not be attracted to those small counties.

4.3 Hierarchical Clustering

We use four different measures on group distance when performing hierarchical clustering, which are:

1. Average: average value of all squared distances of individuals between two groups.
2. Complete: longest distance of individuals between two groups.
3. Median: median value of all distances of individuals between two groups.

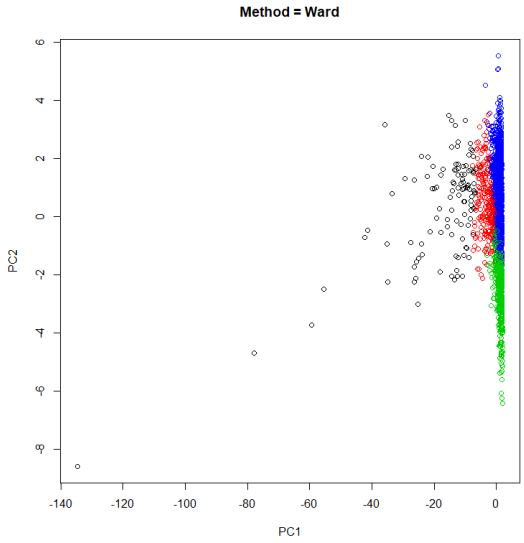


Figure 2: Hierarchical clustering with ward measurement

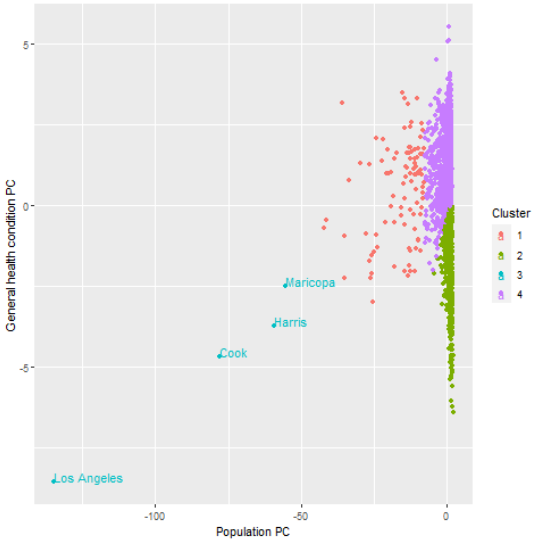


Figure 3: K-means on demographic and health

4. Ward: between group sum of square.

The results are shown in Figure 10 in Appendices. From Figure 10, the first three distance measures lead to exactly the same results if we set the cluster numbers to be 4. The majority of counties fall in the same cluster and the rest of clusters contain very few counties. However, for ward method, it tends to get similar-sized clusters, and therefore has a very different result.

If we further examine clustering results of ward method on top 2 PCs, we can see from Figure 2 that it roughly divide all counties into 3 clusters by their population, which shows by color black, red and blue+green. Furthermore, in the group where the total population is the smallest (color blue+green), the subgroups are divided according to their health related information.

4.4 K-means Clustering

4.4.1 Clustering Analysis

Now, perform the K-means clustering approach on the demographic and health-related information data. From Figure 11 in Appendices and by the elbow method, we choose $K = 4$. The clustering result is shown on Figure 3

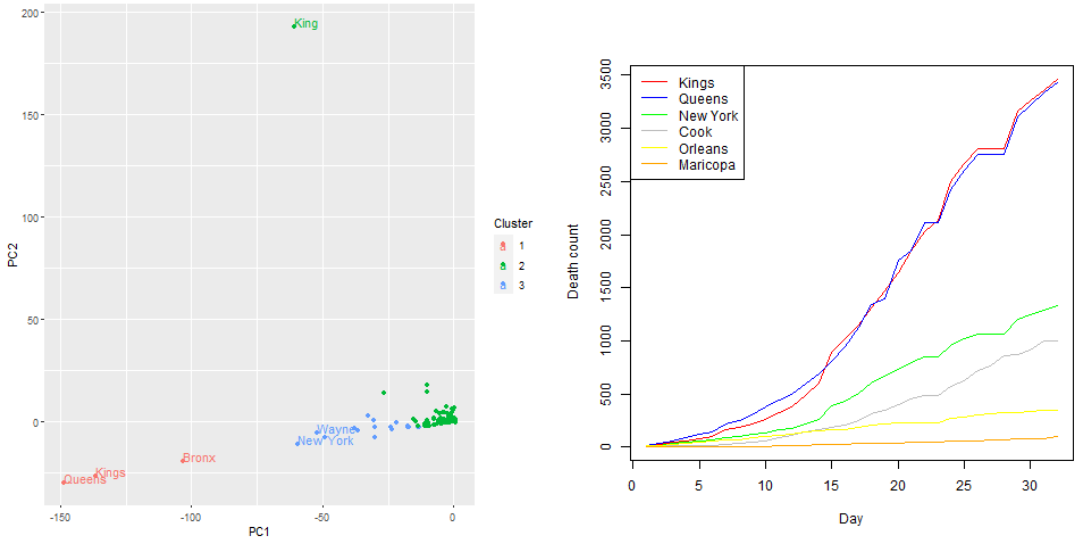
It is not surprisingly that the result is similar to the hierarchical clustering result with ward measurement. Firstly, most of the counties fall into cluster 2 and 4, where they can be labeled as small-sized counties. Again, LA, Cook, Harris and Maricopa are large-sized ones. Secondly, from the second principle component, it can be concluded that small counties have a wide range of health condition due to the reason that huge wealth gap

between the rich and poor counties will have a directly impact on the health condition of the counties.

4.4.2 Pattern Recognition

Now, we apply again the K-means clustering to consider the grow pattern for each county. Here, we only consider the death cases in the analysis.

In order to have a better understanding of the problem, we perform data analysis on the death count for each day. In the first step, we count the total number of day for each county that has death count ≥ 3. We conclude that the maximum number of day is 52, which will be used as time range in this part. Then, in the second step, we calculate the new death cases between two adjacent dates and create a list of variables named "death count". The number of element is the same as the time range 52. Now, it's time to do the K-means clustering. By applying the elbow method again, we choose $K = 3$. The result is shown on Figure 4a.



(a) K-means on death cases

(b) Grow pattern on death cases

Figure 4: Pattern analysis on death cases

To visualize the K-means clustering result in a 2-D dimension, we apply PCA on the data. However, in this case, it is hard and unnecessary to interpret the meaning for each principle component. Thus the discussion on this part will be omitted. From the figure 4a, we know that the result is consistent with the column "Total death cases" in the data-set, where the counties are ordered by the total number of deaths. Then, it is reasonable to conclude that there is relationship between the grow pattern on death and the total number of deaths. Therefore, we pick two counties from each cluster (Kings and Queens from cluster 1, Orleans and Maricopa from cluster 2, New York and Cook from cluster 3) and plot the death count for them w.r.t each day. The result is shown in Figure 4b. From Figure 4b, it can be easily to find out that there are three different

patterns: exponential, linear 1 and linear 2. For exponential pattern, the growth rate for the death is exponentially large, which can be found in county Kings and Queens. For linear 1 pattern, the growth rate for the death is nearly constant but the growing trend can be easily observed, like county New York and Cook. However, for the last one, even if the growth rate is a constant, it is rather slow compared with the pattern linear 1. Then it means that there are not many death cases in this category. The represented counties are Orleans and Maricopa.

4.5 Conclusion

All results of these clustering methods can be summarized into 2 kinds. For k means and hierarchical clustering with ward method, they generally divide the counties into similar-sized clusters based on their population and health facility and condition. For hierarchical with other methods and spectral clustering, they include almost all counties in one cluster and separate them with few special counties like LA and NY. Both results are reasonable, but the first one is more useful for further problems. Besides, it tells us that for medium and large counties, their health level are similar, but small counties seems to have very different health measurement.

5 Supervised Learning

5.1 Classification

5.1.1 Data Processing

In order to remove the effect of total population from the demographic and health related information, we transform the population and health resources amount into rate by dividing the total population number of each county. Also, we scale the data to the same level to perform the classification algorithm. In the classification problem, we divide the data set into training and testing sets by splitting them according to the ratio 7:3. Then, we perform either 10-folds cross validation or other methods to tune the hyper-parameters on the training set. Next, we exam the model performance on the testing set.

5.1.2 Random Forest

We here use random forest to model this categorical response. To tune important parameters of the model, **mtry** and **nodesize**, we use grid search and r package *ranger*, which is 6 times more efficient than the *randomForest* package. Since for random forest, we can use out of bag data to tune parameters, so we don't need cross-validation, or divide data for validation.

After tuning the parameters, we get approximately 0.16 classification error on out-of-bag (OOB) data, which is shown in Figure 12 in Appendices. If we denote the label 1 as positive case, then the green line depicts the error w.r.t. the number of trees for the scenario where prediction indicates label 1 but the true label is 0 instead, which is

a false positive case. Same discussion here for the red line (false negative case). And the black line means the total out-of-bag error. Furthermore, we draw importance plot of variables. As shown from the Figure 13 in Appendices, population density, stay-at-home policy and total population is the 3 most important variables. Besides, some other policies, medical resource, senior people percentage and smokers percentage are also of significant importance. Interestingly, citizens' political stand also seems to play an very important role, but it's hard to determine exactly how it works.

5.1.3 Support Vector Machine (SVM)

It is well-known that it is hard to interpret the SVM result. Therefore, we will only provide the results of tuning parameters and the model accuracy.

There are three hyper-parameters considering to be tuned: **kernel function** (linear, polynomial, radial, sigmoid), **gamma** (parameter needed for all kernels except linear) and **cost** (constant for the regularization term). Table 1 shows the optimal hyper-parameters in the SVM.

kernel function	Gamma	Cost
Linear	0.5	0.1

Table 1: 10-folds cross-validation result for the SVM

Moreover, the overall accuracy on the testing set is about 75.6% and the confusion matrix is shown in Table 2.

		Prediction	
		0	1
Reference	0	518	67
	1	169	213

Table 2: Confusion matrix for SVM

5.1.4 Logistic Regression

We apply the build in function *cv.glmnet* of r package *glmnet* to tune the hyper-parameter for the logistic regression. The optimal tuning parameter **lambda.min** is 0.0024. By applying the Lasso penalty, we obtain the overall accuracy on the testing set is about 73.2% and the confusion matrix is shown in Table 3.

From Table 2 and 3, we know that both classification models have a higher false negative rate (17.5% and 23.2%), which means the model will mis-classify the higher death rate to lower death rate.

		Prediction	
		0	1
Reference	0	549	36
	1	224	158

Table 3: Confusion matrix for logistic regression

5.2 Regression

In order to predict the number of death on April 29, we need to propose some assumptions and construct the basic framework of our models. Since the information after April 22 is assumed to be unknown, we can only use the information before the last day. Therefore, within just one week time period, it is reasonable to assume that the new number of death cases is similar between two adjacent weeks. Therefore, we construct the model by using the deaths and confirmed cases one week from April 22 (which are from April 13rd to April 15th). Besides, all demographic information and health-related variables are considered as predictors. More importantly, the difference between death numbers on April 22 and April 15 is selected as the response variable. Therefore, under the assumption that since April 29 is only one week from April 22, the model structure still holds well. Therefore, we can estimate death numbers roughly on April 29.

5.2.1 Linear Regression with Lasso Penalty

Firstly, we use linear regression model with lasso penalty to model the difference in death. Furthermore, we use the nearest 5 day of cases and 2 days of deaths one week from April 22. This is selected with cross-validation, and the lasso penalty parameter is tuned by the build in function from *glmnet* package.

It is not surprisingly that information on deaths and cases which are close to the closing date are most important predictors. Besides, we care more about other variables such as demographic and health-related information, which will provide moe useful information for further discussion. In other words, we need to pay attention to the estimated parameters from demographic and health-related variables that are more significantly non-zero from the Lasso penalty. Figure 5 shows the results of non-zero age related estimators comparison between male and female among all the age groups. Clearly, female are more vulnerable to COVID-19 than male, and older people seems to catch the disease easier than younger ones. Moreover, better health resources can help reduce fatality. It's interesting that policies aren't very important according to our model, and the reason behind would be that the policies are announced in a rather similar time period among all the counties, which may not have clear effect difference, and second, some counties with less stress with the virus tend to have the policy later, which kind of cancels out the effect policies brought to the more serious region.

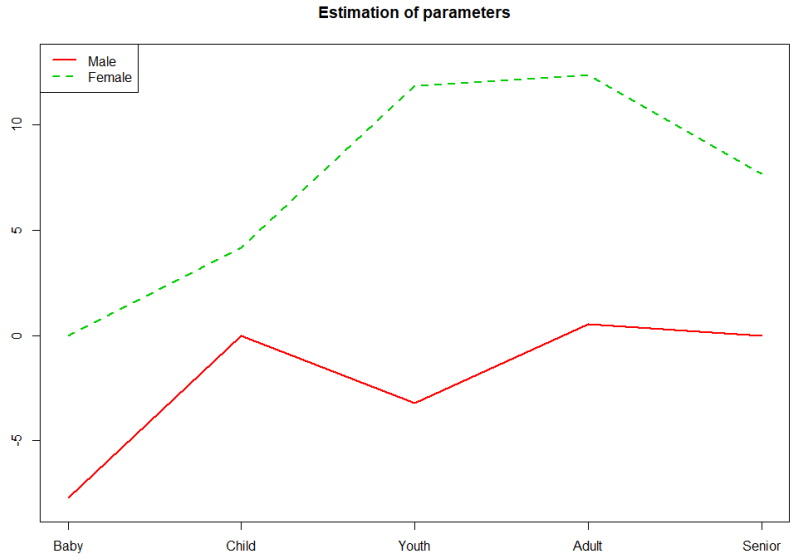


Figure 5: Non-zero age related estimators comparison between male and female among all the age groups

5.2.2 Random Forest

We again use random forest for regression problem. Parameter tuning and model building procedures are similar to classification problem. Unlike the data being sued in linear regression with Lasso penalty, we only consider the deaths and cases information on the latest day that is one week from the closing date instead of tuning the optimal number of days by cross-validation. Since random forest is a black box algorithm, we can only interpret the result by analyzing the variable importance result. The result is shown in Figure 14 in Appendices.

The result is similar to linear regression with Lasso penalty. Amount of deaths and confirmed cases from the latest date are most important among all the others. Female population has much more influence on the prediction. Age group can also show an impact. Besides, some of the health related information such as number of ICUbeds, number of full-time employees at hospitals in 2017 are also important variables. However, random forest stresses more weights on policies comparing with regression with Lasso penalty, which may due to the interaction effect included in the random forest model.

5.2.3 Boosting

Next, we perform a boosting algorithm with xgbTree method in this section. Firstly, we select again all the demographic and health related information. Besides, we include the total number of deaths and confirmed cases one day before the closing date, that is the date April 15th. Different from the data processing in the previous modeling, there is no need to scale the data into the same magnitude because the boosting algorithm can take

care of the scaling issue. However, hyper-parameters are still needed to be tuned by 10-folds cross validation. The main tuning parameters are: **number of trees (nrounds)**, **number of splits (max_depth)**, **learning rate (eta)**. The tuning result is shown in Table 4.

Parameter	nrounds	max_depth	eta
Value	50	3	0.4

Table 4: 10-folds cross-validation result for the Boosting

Besides, by similar discussion in the random forest, we also point out the most important variables in the boosting algorithm, which is shown in Figure 15 and 16 in Appendices.

To show the model prediction results, using the same way as presented in the reference paper, we choose four counties to compare the prediction result: Queens, Cook, New York and Orleans. The comparison is shown in Table 5.

County name	Total death April.22	Total predicted April.29	Total actual April.29
Los Angeles	544	842	1056
Cook	1002	1409	1516
Orleans	344	424	416
Wayne	1278	1687	1727

Table 5: Prediction result in boosting

It can concluded that the performance of boosting algorithm is fine in county Orleans and Wayne, but not as good as expected in county Los Angeles and Cook. The reason behind it is the spread of the virus and the corresponding death rate will be influenced by more complicated reasons in big counties such as Los Angeles and Cook.

5.2.4 Bonus

We acquired newly updated information and use it to validate our Lasso regression model. From predicted v.s observed plot in Figure 6, the model serves good for most of counties except those with large death numbers, mainly Bronx, Queens and Kings in New York. For these 3 counties, prediction seems to underestimate the deaths increase a lot, which indicates that they shared a growth pattern different from other counties. This result is supported by our growth pattern study in unsupervised learning section, where these 3 counties are accurately categorized into the same cluster and their death number growth shows an exponential feature.

Since exponential pattern is only shown in these 3 counties, we may consider adding exponential terms to model their pattern. But since we only have 3 counties with this growth pattern, we have to utilize the information of previous days to train the parameters. This idea is similar to the method proposed in Altieri et al. (2020), except we only assume non-linear effect on increment of deaths of these 3 counties to keep results more robust.

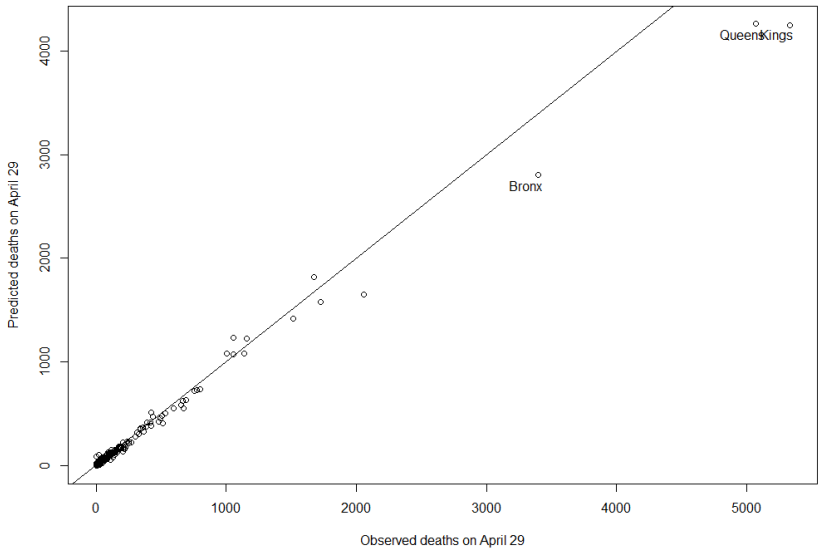


Figure 6: Comparison between the prediction and actual death cases on April 29th

6 Collaborate Question and Concluding Remarks

As presented before, we divide the residents in to five categories based on their age, along with other variables that may reflect residents' profile in a county. Then in previous sections, we use these variables to build several models to deal with classification and regression problems related to the severity of the disease. Intuitively, if the model suggests that a large number of one resident group may cause the pandemic to be more serious, we can conclude that the group is more vulnerable to the virus. The idea is similar when it comes to medical resources and policies.

The answer to identifying vulnerable sub-population mainly comes from linear regression with Lasso penalty. The estimated parameters are comparable to each other since we scale the data before fitting the model. We then conclude that the main result with regard to resident profile from Lasso regression, which contains estimated parameters for different sex in different age stage. It shows that a male clearly has more resistance to the disease than a female, and the older ones are more vulnerable than the younger ones.

From the random forest model for classification, it seems that population density is the most important factor. If we further investigate how it works, we get the result shown in Figure 7.

From Figure 7, all counties having more than 1 death per 100000 people have a population density more than 5000 per square mile. So residents from a more crowded county have higher probability of dying from this virus, which may due to better chance to get infected and less available medical help instead of themselves.

Moreover, from Lasso model and Table 6, medicare enrollment and number of ICU beds are also important variables, which kind of reflects two aspects of dealing with infected

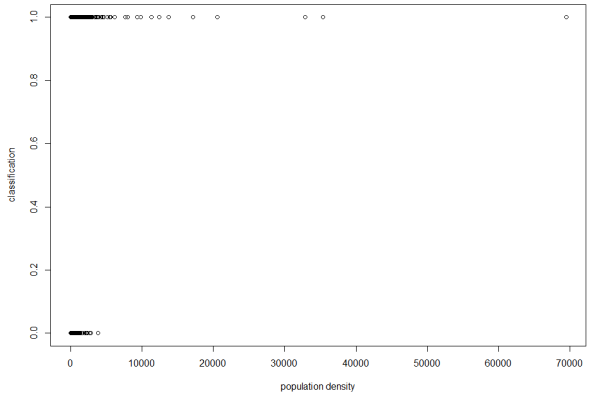


Figure 7: Population density among all the counties

residents: they have to have financial ability and the hospital need to have available resources to make the patient receive proper medical treatment.

Covariate	Estimation
PopTotalMale2017	-17.949878064
MedicareEnrollment. AgedTot2017	-5.686152209
ICU_beds	-4.911475862

Table 6: non-zero estimation of the variables in linear regression with Lasso penalty

Based on those findings about the disease, we propose three feasible measures to help reduce mortality:

1. Protect the vulnerable population. Government should pay more attention to nursing home where people lives in a high density and are more vulnerable to this disease.
2. Make sure patients get proper treatment. Government should offer medical subsidies for those who can't get medical treatment because of their financial problem. Also, temporary hospitals should be built to help deal with the emergency.
3. Reduce population density. It's impossible to literally reduce density, but those policies are all actually used to prevent large group gathering, which is indeed a way to reduce density.

Most of the anti-contagion policies don't express much influence in our models, and we have already briefly talked about the reasons: small time range for effect, low variation and cancel-out effect. Some of them are supported by the results in Hsiang et al. (2020).

Appendices

Appendix A Details in Data Processing

1. Deleting the whole column, there are several cases:

   (a) Useless information: first column and the columns where there are no cases and deaths report for all the counties.

   (b) Replaced by the other column that contains full values: replace "lat" and "lon" by column "pop-latitude" and "pop-longitude"; replace column "3-year diabetes" by column "diabetes percentage";replace "fracfemal" by "poptotalfemale"; replace "age65+2017" by group "Senior" (which is defined below).

   (c) Too many missing values: columns "3-year mortality" for all ages and column "3 year total mortality".

2. Imputing the missing values by the mean of the column if the number of missing values is less than 5%. For example, column "medical enrollment", "eligible for medicare", "diabetes percentage", "heart disease mortality", "strike mortality", "dem to rep ratio" and "SVI percentage".

3. Imputing the missing values by some policy-related assumption. For the missing values in columns "stay at home", "> 50 gathering", "> 500 gathering" and "entertainment/gym", it could be the case that the governments of those counties do not have the anti-contagion policy at the time Apr.22, 2020. Therefore, it is reasonable to assume that the missing values could be filled by number 737538 (it is calculated by the "proleptic Gregorian ordinal of the date" from the first day to Apr.22, 2020, where first day has a value of 1).

4. To deal with moderate missing rate variables, HPSAShortage, HPSAServed and Underserved Pop, we use another data imputation method. Noticing that these 3 variables miss together, multiple imputation is not good here. So we simply use all the rest variables related to medical aspect and perform a linear regression and use predictions to replace missing values of those variables.

Appendix B Figures for previous Sections

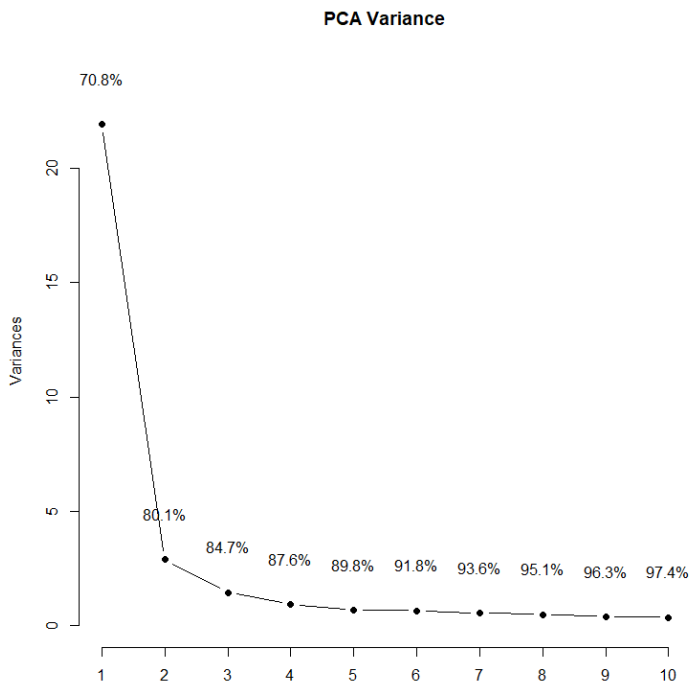


Figure 8: Principle component analysis on demographic and health

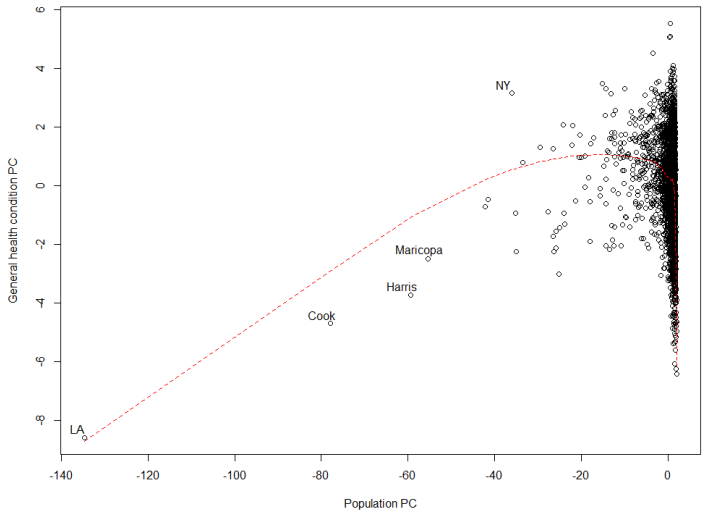


Figure 9: Spectral clustering on demographic and health

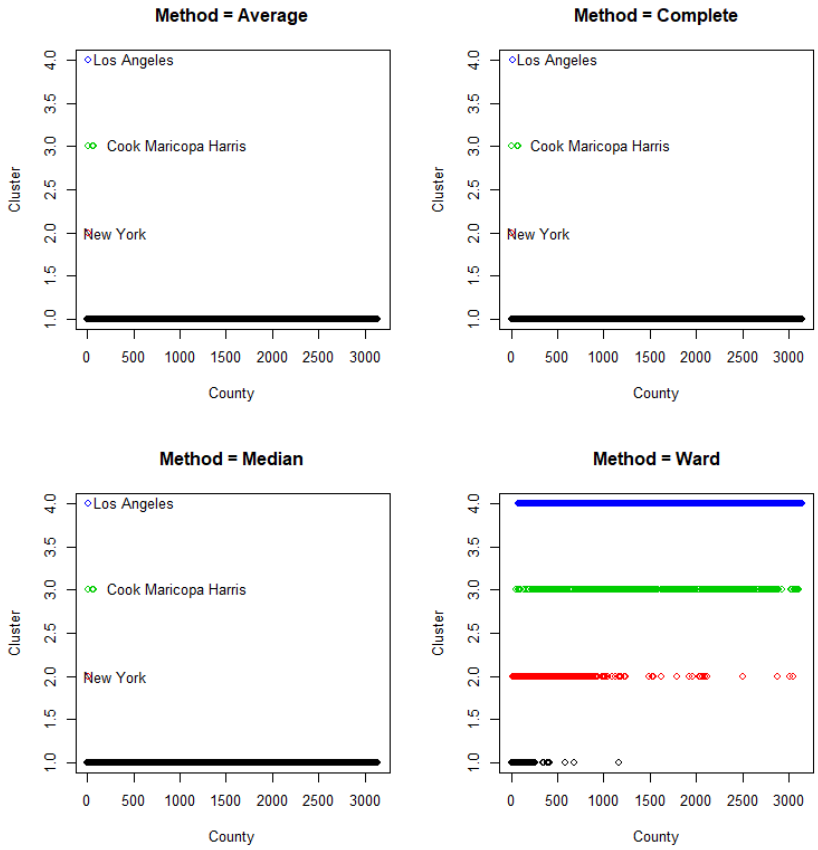


Figure 10: Hierarchical clustering on demographic and health

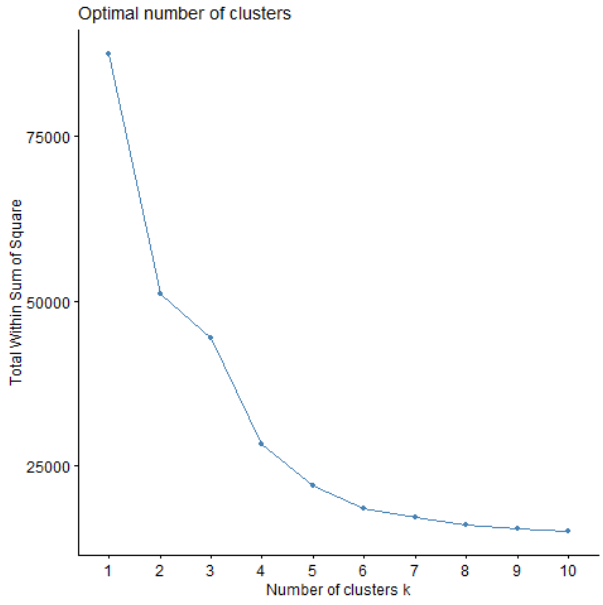


Figure 11: Elbow method to choose the number of clusters

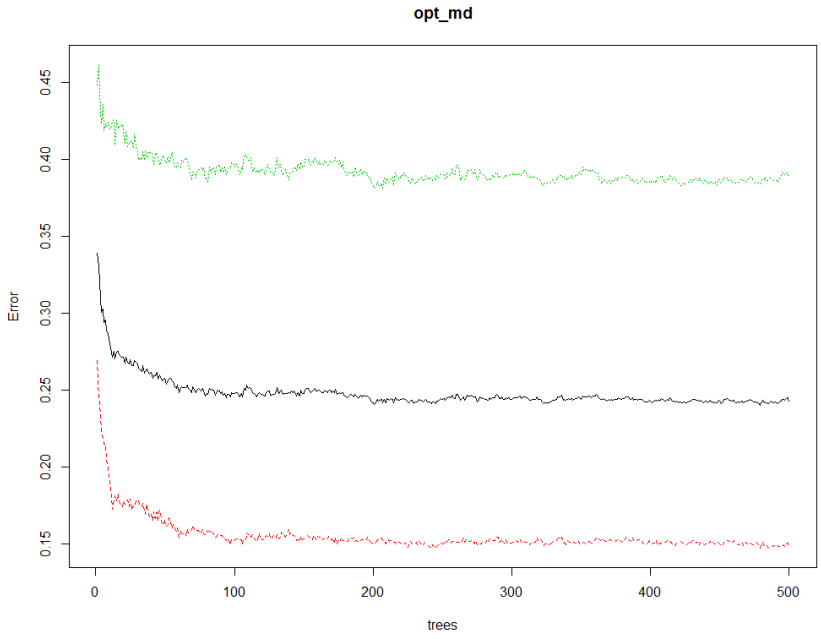


Figure 12: Out-of-bag error w.r.t. number of trees in random forest

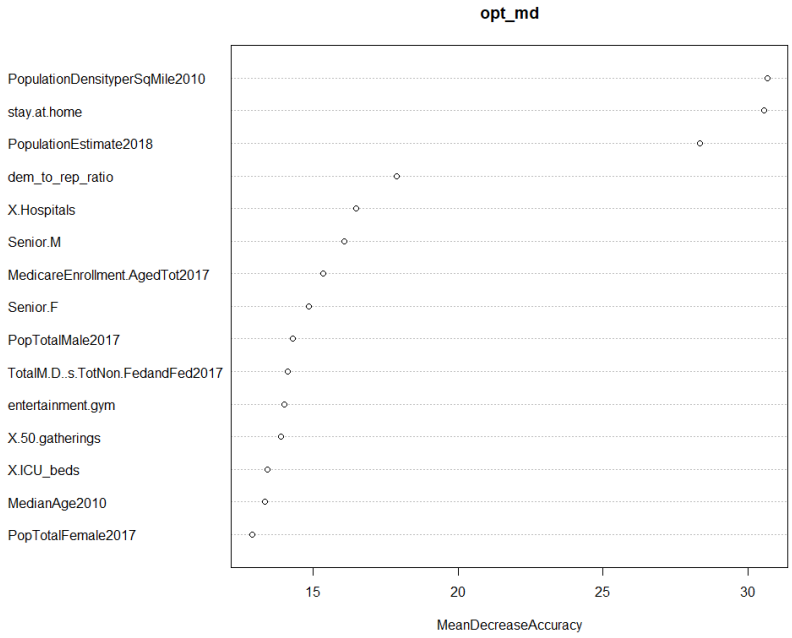


Figure 13: Importance of the variables in random forest classification model

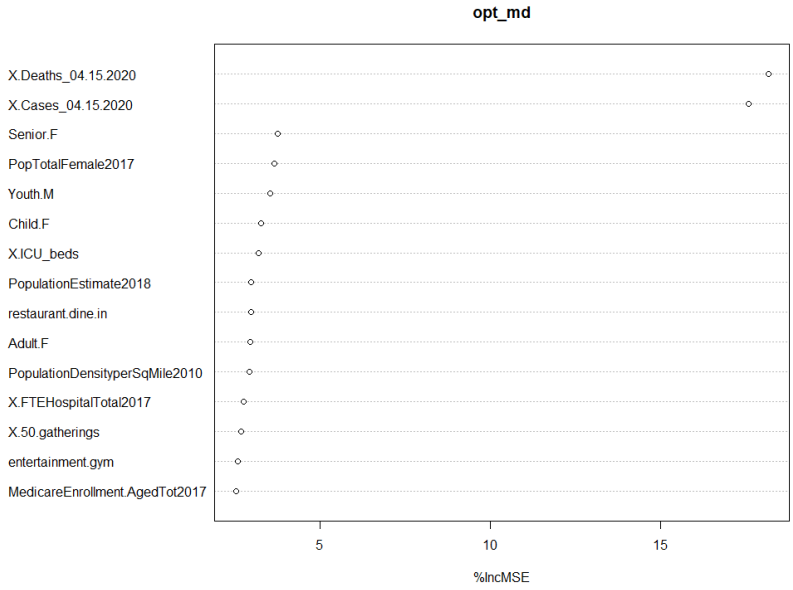


Figure 14: Importance of the variables in random forest regression model

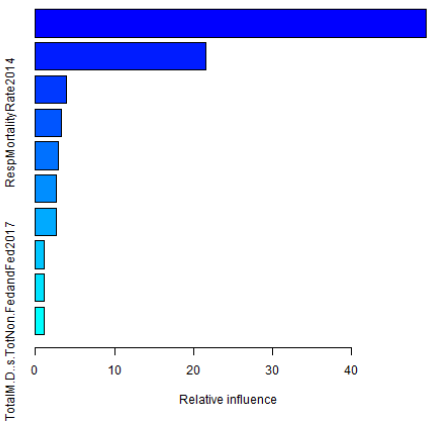


Figure 15: Relative influence of the variables in Boosting

Variable	rel.inf
Confirm	49.47
Death	21.5
PopDenSqMile2010	3.97
MortalityRate2014	3.24
MedicareEnroll2017	2.85
Stroke	2.70
Senior.F	2.67
EligibleforMedicare	1.16
Youth.F	1.14
TotNon.FedandFed2017	1.13

Figure 16: Relative influence

References

Altieri, N., Barter, R., Duncan, J., Dwivedi, R., Kumbier, K., Li, X., ... others (2020). Curating a COVID-19 data repository and forecasting county-level death counts in the united states.

Elmousalami, H. H., & Hassanien, A. E. (2020). Day level forecasting for Coronavirus Disease (COVID-19) spread: analysis, modeling and recommendations. *arXiv preprint arXiv:2003.07778*.

Hsiang, S., Allen, D., Annan-Phan, S., Bell, K., Bolliger, I., Chong, T., ... others (2020). The effect of large-scale anti-contagion policies on the coronavirus (covid-19) pandemic. *MedRxiv*.

Schuller, G. D., Yu, B., Huang, D., & Edler, B. (2002). Perceptual audio coding using adaptive pre-and post-filters and lossless compression. *IEEE Transactions on Speech and Audio Processing*, *10*(6), 379–390.