University of Illinois Urbana-Champaign
STAT 542

FINAL PROJECT:

COVID-19 County Level Data Analysis

Team Lead: Chenghao DING (cd7)
Team Member: Changyue HU (ch47)
Team Member: Jingbin CAO (jingbin2)

2020-5-15

STAT 542: COVID-19 County Level Data Analysis

Spring 2020, by Changyue Hu (ch47), Chenghao Ding(cd7) and Jingbin Cao(jingbin2)

May 15, 2020

Contents

Project Description

This project uses county level data about demographics and health-related information to predict one week deaths from April 23rd to April 30th. We consider some predictor variables, such as the polulation, age distribution, gender distribution, per capita share of medical resources and health conditions. At first, to give a simple understanding, we perform the covid-19 growing pattern clustering by k-means method, demographics clustering by spectral method and health-related information clustering by agglomerative hierarchical clustering method. We use logistic regression and suport vector machine to predict whether death per 100,000 population in the county is larger than 1. Then, four regression methods, elastic penalized regression, random forest regression, GAM regression and XGboost method, are used to to predict one week deaths, based on the previous days' amount of deaths, the cases a week ago and the characteristics of demogrphics and health-related information.

According to the clusters, we find cases and health resource are highly affected by the economic situation of a county. Comparing to them, deaths and demographics do not show any obvious regional characteristics and scatter throughout the country. The classification performs well and can give identify most counties with high death rate.

According to the coefficients of the regression model, we conclude that the the population and population density makes their death toll rise even faster in the county. People over 85 with heart disease are more vulnerable to this virus. Comparing to male, female are more vulnerable. It is helpful for reducing the deaths to provide adequate medical resource, such as a hospital. However, because of the population base, age density and gender density are related to social situation in this area, we think the analysis of vulnerable poeple are affected.

Litureture review

In "Curating a COVID-19Curating a COVID-19 data repository and forecasting county-level death counts in the United States" (2020), the authors (Altieri1 et al.) collate a large database about COVID-19 information and develop several models to forecast short-term deaths at the county-level in the United States resulting from COVID-19.

The database they build includes county level information (such as demographic information and health resource) and hospital level data. When building prediction models, several predictor variables are considered, including the amount of deaths and cases in previous days in the current county and in neighboring counties, a set of demographic and healthcare-related features and so on. For the individual county-specific prediction models, the exponential predictors (Poisson generalized linear model) and linear predictors are both developed. They combine their forecasts using ensembling techniques and get the Combined Linear and Exponential Predictors (CLEP). They think the prediction of the expected number of deaths over the next week will be helpful for county-specific decision-making and give a sense of the future .

Data Preprocessing

Missing Value Treatment

The data set has nearly 276 features, however, there are some features of county containing a lot of missing values. Below we present the variables which have missing values.

```
##   3-YrMortalityAge1-4Years2015-17  3-YrMortalityAge5-14Years2015-17
##                              3076                              3046
##         mortality2015-17Estimated     3-YrMortalityAge<1Year2015-17
##                              3046                              2673
## 3-YrMortalityAge15-24Years2015-17 3-YrMortalityAge25-34Years2015-17
##                              2515                              2184
## 3-YrMortalityAge35-44Years2015-17              3-YrDiabetes2015-17
##                              1843                              1720
```

We first remove the features which contains a large fraction of missing values. In detail, we dropped 25 features which have more than 50 missing values.

In the process of dealing with missing value problem, we notice that there are 5 features that only contains a few missing values (no more than 21): `MedicareEnrollment,AgedTot2017`, `StrokeMortality`, `#EligibleforMedicare2018`, `HeartDiseaseMortality` and `DiabetesPercentage`. We think these county level features are important health resource and risk factors and might be helpful for our analysis and prediction modeling. Given that the number of missing values in these variables is limited, we compensate for missing values of these 5 variables.

Assume that the counties which have the cloest population centers would enjoy similar demograchic and healthe related features. We implement the data imputation procedure for these 5 variables by replacing the missing values with 1-NN estimate, determining the neighbor of a county based on `POP_LATITUDE` and `POP_LONGITUDE`.

Variable transformation

- Age division

According to the age division method in CDC COVID-19 weekly report, we group the county level population by 5 age groups, `<5yrs,5-19yrs,20-64yrs,65-84yrs,>85yrs` and then divide them by the total population of the county to get the estimated propotions of different age group for each county. In detail, we create five new variables `%pop<5`, `%pop5-19`, `%pop20-64`, `%pop65-84` and `%pop>85`, which are approximate percentage of total population for these five age groups.

- Per capita level health resource

Considering that the health resource of a county, like `#Hospitals` and `#ICU_beds`, is correlated with the total population of the county, we transform health resource information on per capita level to get a more accurate acknowledge of a county's medical condition.

Clustering

In order to understand the data, especially the underlying COVID19 counts pattern at the county level and its association with demographics and health-related information of a county, we will perform three clustering in this section. In detail, we will perform the covid-19 growing pattern clustering by k-means method, demographics clustering by spectral method and health-related information clustering by agglomerative hierarchical clustering method.

Case and Death Growing Pattern (Kmeans Cluster)

To find the underlying pattern (at the county level) in terms of how the COVID19 counts are growing and the death counts are growing, we use two variables (r_k and r'_k) as explanatory variables to cluster the counties.

For both cases and deaths, we calculate the variables as following.

Set growth rate r_{ki} for county k in day i is

$$r_{k,i} = \frac{\sharp \text{ of case}_{k,i+1} - \sharp \text{ of case}_{k,i}}{\sharp \text{ of case}_{k,i}}$$

The average growth rate r_k for county k is defind as:

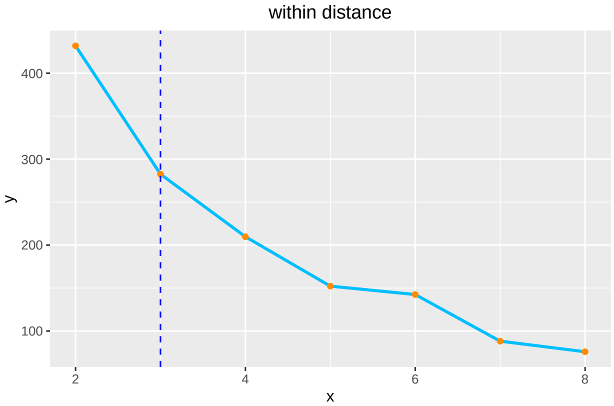
$$r_k = \frac{1}{\sharp \text{ of days}_k - 1} \sum_{i=1}^{\sharp \text{ of days}_k - 1} r_{k,i}$$

Which is similar to the average of derivatives of case county.

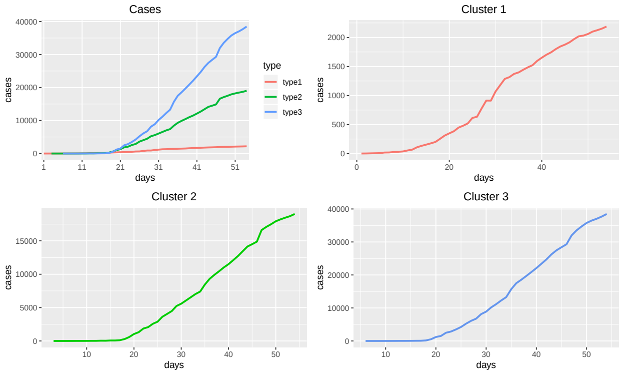
r'_k is similar to the average of the second derivatives of case county:

$$r'_k = \frac{1}{\sharp \text{ of days}_k - 2} \sum_{i=1}^{\sharp \text{ of days}_k - 2} \frac{r_{k,i+1} - r_{k,i}}{r_{k,i}}$$

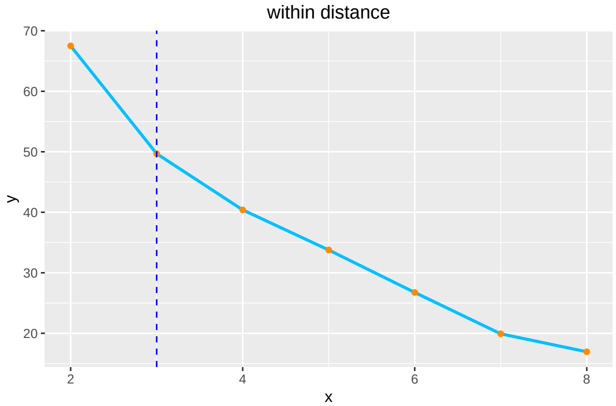
Cases:



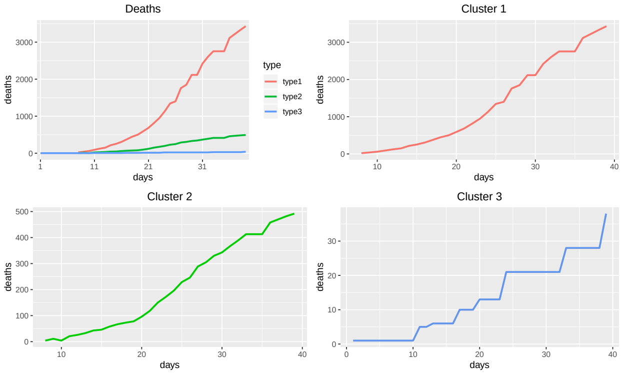
In the above figure, we see the with-in distance has a obvious inflexion when the number of cluster is 3. Therefore, we find 3 clusters.

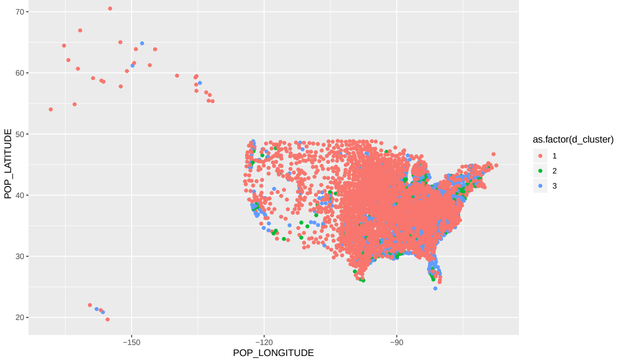


Death:

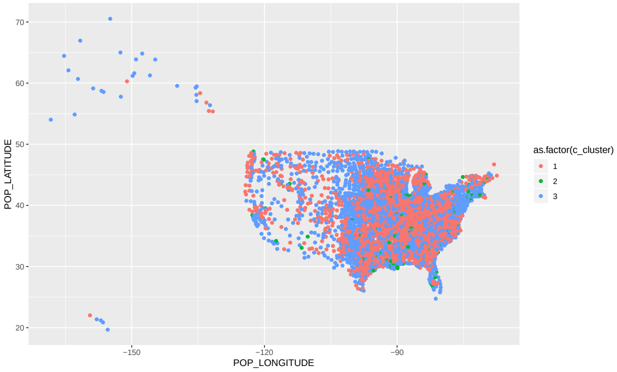


Similarly, we choose 3 clusters:





In the above figure, it shows the clusters for cases growth pattern in U.S. Countyies in the First cluster are shown in red, counties in the second cluster are shown in green and counties in the third cluster are shown in blue. It shows that most counties have different pattern with the counties in the southwest and eastnorth. This result is consistent with what we now know. The counties in New York, Log Angeles areas and Florida grow fastest, the counties near them also grow faster than others.

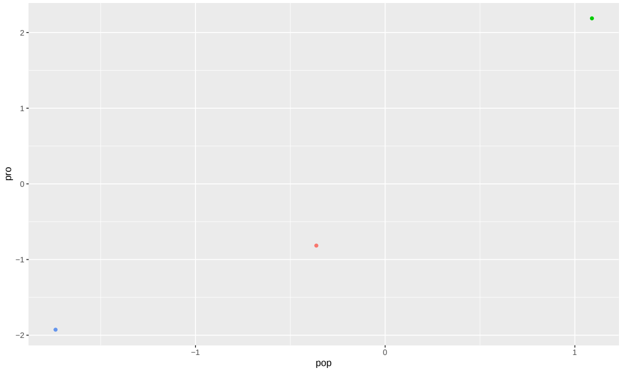


This Figure shows the clusters for deaths growth pattern in U.S. It is very different from the previous figure, which means the deaths pattern are not similar to cases. The cases growth fast is not necessary to have higher deaths increasing rate. It may casued by the demographics and medical condition. We will try to explore it in the following sections.

Demographical Information

We use the near 5 neighbors as the adjacency matrix. We choose to have 3 clusters to compare with pattern.

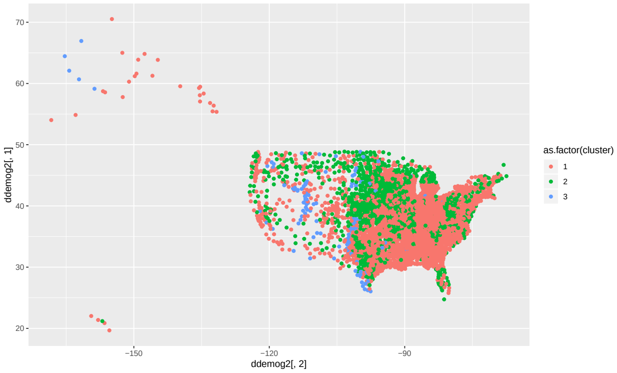
For each cluster, we show the center of population and the proportion of people over 65.



The three clusters have obvious different characters in demograhics.

The cluster result for demographics are shown in map.

```
##
##    1    2    3
## 2125  897  119
```



In the figure, we can find that the some economically developed areas have a similar population characteristic (they are in the same cluster). Also, the counties in each cluster are not concentrated, but scattered throughout the country, which may contribute to the prediction of deaths increase.

Health-related Informaton

To explore the difference of health-related condition for different counties, we perform hierarchical clustering based on the health-related information. We want to see if there are any underlying clusters and if the result is similar to the COVID-19 pattern clusters.

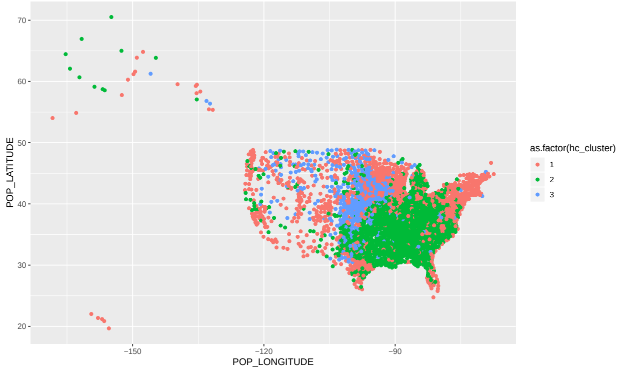
Among the hierarchical clustering methods, we choose agglomerative HC. Agglomerative clustering is also known as AGNES (Agglomerative Nesting). It works in a bottom-up manner. That is, each object is initially considered as a single-element cluster (leaf). At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster (nodes).

We perform agglomerative HC with `agnes`. With the `agnes` function we can also get the agglomerative coefficient (ac), which measures the amount of clustering structure found (values closer to 1 suggest strong clustering structure). This allows us to find certain hierarchical clustering methods that can identify stronger clustering structures. Here we see that Ward's method identifies the strongest clustering structure of the four methods assessed.

Based on he result of agglomerative HC with Ward's method, we group counties into 3 clusters, which is consistent with the number of clusters in COID-19 patterns. The table below presents the number of counties in each cluster.

```
## hc_cluster
##    1    2    3
## 1268 1465  408
```

To visualized our clustering result, we plot the three clusters of counties on a demographical map like before.



For the plot above we see the counties in the first cluster are mainly distributed in states like Washington, California, New York and Florida, which are all states with good economic conditions. The clustering result based on health-related information seems related to economic condition of the counties. comparing this result with the COVID-19 pattern clustering result, we see that the relationship between a county's medical resource level and its covid-19 situation could not be linear. The potential association between number of covid-19 cases and the healthe resource would be much more complex. Higher levels of medical resources and care are usually found in large cities. But on the other hand, the high population density and movement of people in large cities often accelerate the spread of the virus, making large cities, such as New York, the epicenter of COVID-19.

Classification

In this section, two methods such as Support-vector machine (SVM) and Logistic regression are applied to this binary classification problem. First of all, define the class variable-Death per 100,000 population > 1, and label the repsonse variable in training data accordingly. Considering this is a complicated case with a non-seperable boundary, therefore, a nonlinear SVM with a radial kernel is used to fit the train data. Also, hyperparameters such as cost of constraints violation and gamma are tuned using tune.svm function and searched by grid. Finally, the best cost and gamma parameter are found. The searching grids of cost and gamma are [0.1,1,10]. Before classification, the train data are scaled.

```
##        target
## pred3   FALSE TRUE
##   FALSE  1692  435
##   TRUE    219  795
```

Thus, the best tunning Cost=1, gamma = 0.1. The prediction accuracy is 0.7917861.

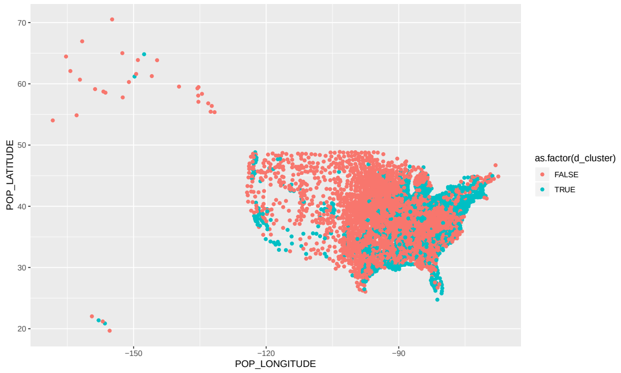
Second method is Penalized logistic regression. A ten-fold cross validation is applied. Penalized linear regression (with lasso penalty) is implemented.

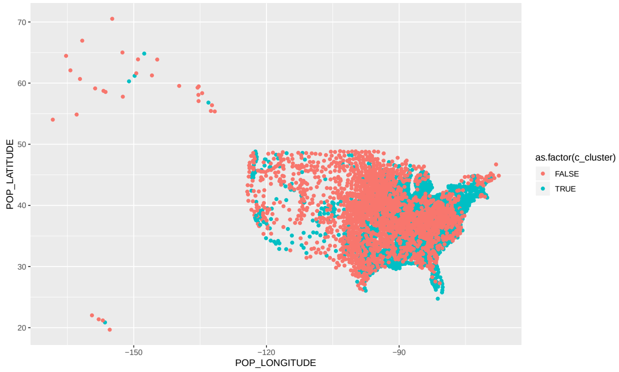
```
## 15 x 1 sparse Matrix of class "dgCMatrix"
##                                         s0
## (Intercept)                    -0.48639020
## PopulationDensityperSqMile2010  2.36706527
## MedianAge2010                   0.14528630
## DiabetesPercentage              0.16773810
## HeartDiseaseMortality           0.16160186
## StrokeMortality                 .
## Smokers_Percentage             -0.05186802
## RespMortalityRate2014          -0.24579965
## X.Hospitals                    -0.80221516
## X.ICU_beds                      0.36195500
## X.pop.5                        -0.18318110
## X.pop5.19                       0.46853870
## X.pop20.64                      0.31163039
## X.pop65.84                     -0.15622086
## X.pop.85                       -0.06556930
```

The regression coefficients is shown above too. The population Density has the largest importance in prediction.

The best λ=7.7088932 × 10^{-4} and prediction accuracy on training data is 0.6625279.

Finally, the classfication results using SVM is compared with the true value is the following US map.





It is found that the the main difference between this two plots are the classification in the middle part of USA. It failed to classify the death ratio in central US. However, for the other parts of USA, especially the most serious districts such as New York, Seattle, Texas, and Chicago are correcly classified. This is reasonable consider there are a clear different death growing rate if you compare rural areas or less developed towns with big cities.

Regression: Death Count Prediction

Model

In this section, we mainly use four method to predict the amount of deaths at county level in the following week (from April 23rd to April 30th).

The relationship between respond variable (amount of deaths in day t in each county) and predictor variables:

$$\log death_t = \beta_0 + \beta_1 \log death_{t-1} + \beta_2 \log case_{t-7} + \sum_{k=3}^{21} \beta_k x_k \tag{1}$$

The x_k's are the variables about demographics and health related information at county level introduced in *Section “Data Processing"*.

Training data set

We use the data before April 23th to train the model. Because of the “log" form, we remove the days with deaths= 0. Also, we think when the deaths is equal to 0, it may have different pattern (stay for a few days).

Elastic Linear Regression and Stepwise Method

We want to use the simple linear regression(OLS) to build the model, but we prefer less variables. Thus, consider Elastic Penalty (with $\alpha = 0.5$) to reduce the dimention at first.

Elastic penalized method:

It is a method combine the l_1 penalty and l_2 penalty. For a given λ_1, λ_2 and loss function$L$:

$$L_{Elastic}(\beta) = L(\beta) + \lambda_1 \sum_{k=2}^{p} |\beta_k| + \lambda_2 \sum_{k=1}^{p} \beta_k^2, \qquad \hat{\beta}_{Elastic} = \arg\min_{\beta} L_{Elastic}(\beta) \tag{2}$$

where $L(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\beta))^2$. When $\alpha = 0.5$, λ_1 is equal to λ_2.

After this selection, use stepwise algorithm and AIC to check other variables.

The Akaike information criterion (AIC) is an estimator of out-of-sample prediction error. AIC provides a means for model selection. In estimating the amount of information lost by a model, AIC deals with the trade-off between the goodness of fit of the model and the simplicity of the model.

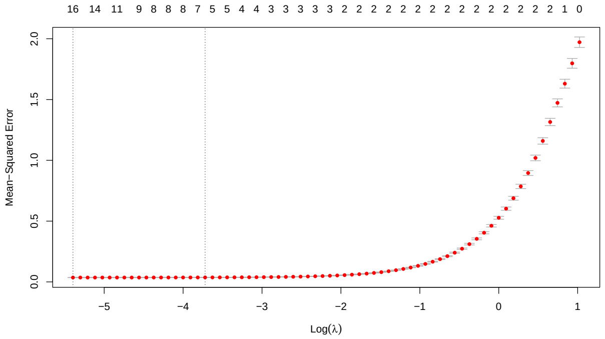
$$AIC = 2(-L(\hat{\theta}) + (k+1)), \tag{3}$$

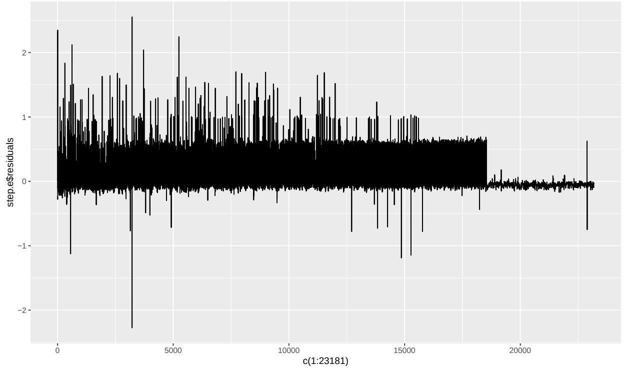
where L is the log-likelihood funtion.

AIC compensate for the number of parameters to get simple model. Stepwise method can choose the model with smallest value of AIC.

Stepwise: use different combinations of variables to build model and select the combination with smallest AIC.

Completing the above two steps, rebuild simple linear regression model use the selected variables, and we get the results in the following table.





According to the residuals plot, the model fits better when deaths is large.

Generalized Additive Model

GAM (Generalized Additive Model) is a powerful and yet simple technique for regression problem. It has substantially more flexibility than GLM because the relationships between independent and dependent variable are not assumed to be linear. In fact, we don't have to know a priori what type of predictive functions we will eventually need to predict the future death count. From an estimation standpoint, the use of regularized, nonparametric functions avoids the pitfalls of dealing with higher order polynomial terms in linear models. We try to predict $logdeath_{t-1}$, given $logdeath_{t-1}$, $logcase_{t-7}$ and other explanatory variables by fitting a GAM model. Formula has been given below in the R code chunk.

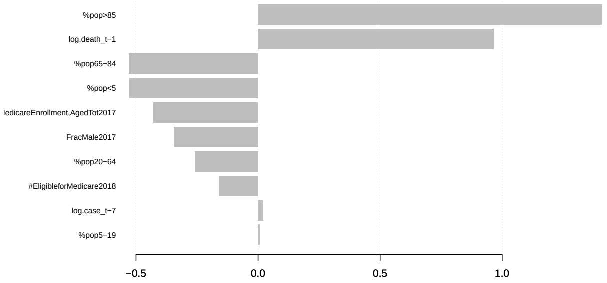
Boosting

In this part, a Gradient Boosting method using xgboost is used to predict the next week death toll. 10 fold cross-validation is used to find the best tuning paramter. A linear weak learner is forced to learn the train data. Sum of squared residuals are used as the evaluation metric. A grid of search of hyperparamters in gradient boosting is conducted. The hyperparameters fitted are max_depth of built trees, learning rate.

Table 1:

	Dependent variable:	
	stepwise	
log.death_t-1‘	0.967***	(0.002)
‘log.case_t-7‘	0.022***	(0.001)
FracMale2017	−0.487***	(0.083)
PopulationEstimate65.2017	1.719e-07***	(0.00000)
PopulationDensityperSqMile2010	$1.324e - 06$ ***	(0.00000)
MedicareEnrollment.AgedTot2017	−0.409***	(0.058)
HeartDiseaseMortality	0.0003***	(0.00004)
StrokeMortality	−0.001***	(0.0002)
RespMortalityRate2014	$-5.864e{-}04$***	(0.0001)
TotalM.D..s.TotNon.FedandFed2017	$1.239e{-}05$*	(0.00001)
X.HospParticipatinginNetwork2017	−0.003***	(0.001)
X.Hospitals	−0.002***	(0.001)
X.pop.5	−0.619***	(0.174)
X.pop.85	1.660***	(0.303)
Constant	0.365***	(0.052)
Observations	23,181	
R^2	0.982	
Adjusted R^2	0.982	
Residual Std. Error	0.189 (df = 23166)	
F Statistic	90,253.070*** (df = 14; 23166)	
Akaike Inf. Crit.	-11541.82	
Note:		*p<0.1; **p<0.05; ***p<0.01

Considering there are three different clusters of counties that we found before. Also, the best tree depth and learning rate are tunned separately in these three models. The root mean square error are evaluated to find the best tuning parameter. Also, the feature importance is plotted.



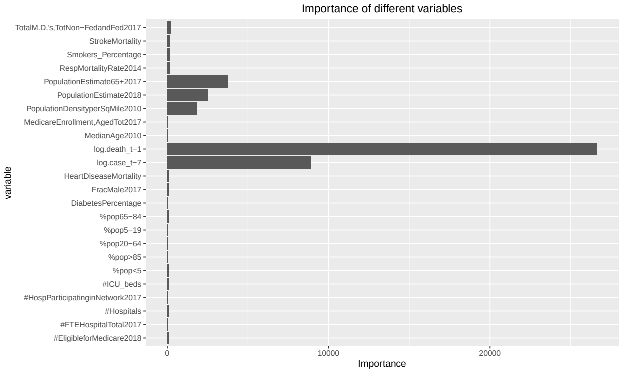
The best tunning paramter of learning rate is 0.2932461, and max_depth is 10.

The mean sqaured error of prediction in training dataset will be 0.0354181. The most important factor that are affecting death toll is the proportion of senior people of age older than 85.

The second most important factor that are affecting prediction of death toll is the death number one day before in that county. This is supported by the COV-19 pandemic study of WHO that senior people are more vulnerable to COV-19 disease.

Random Forest

Random Forest constructs a multitude of decision trees and outputs the mean prediction regression of the individual trees. Each tree are built based on a random sample with replacement of the training set. When building a tree, each time it consider one rule based on one dimention and use the rule to split the data into two parts. After several splits, there are many parts each of which has many samples. The mean of the samples in one part are the prediction of a prediction point if the point falls into this part according the rules we build.



In the random forest method, according to the rules we build in each tree, we get the importance for each predictor variable shown in the above figure. According to this, we find the population have important effect.

The table below presents the training MSE and RMSE for each method. We see that the Generalized Addictive Model has the smallest MSE among the four methods.

	Elastic & Stepwise	GAM	Random Forest	XGBoost
MSE	0.03554	0.03458	0.03976	0.03542
RMSE	0.1885	0.1860	0.1994	0.1882

Table 2:

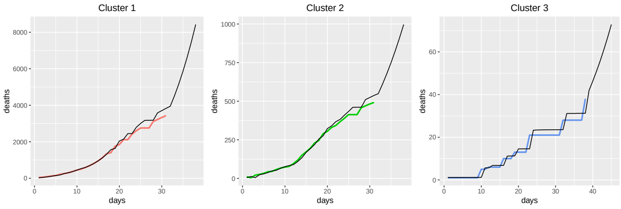
Prediction data set

We build prediction matrix for counties shown in deaths clusters and use our four models above to predict the number of death one week from Apr 22.

Prediction Result

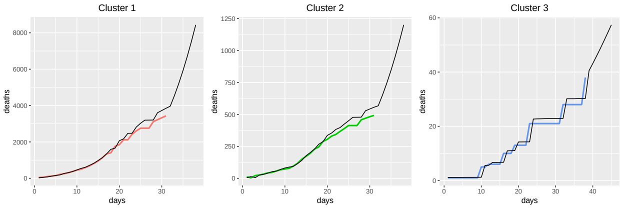
Prediction: Stepwise model

Plot



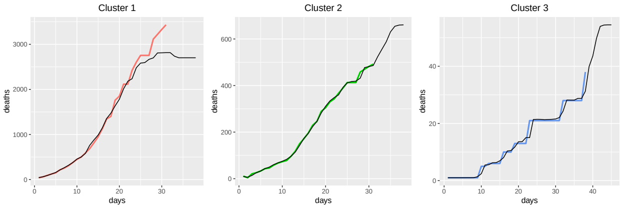
Prediction: GAM model

Plot



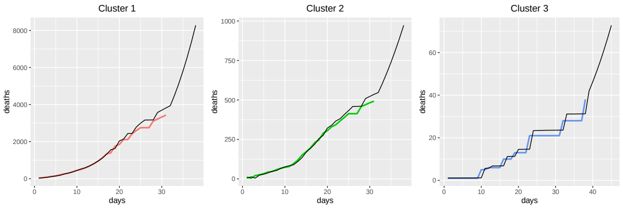
Prediction: Random Forest model.

Plot



Prediction: Gradient Boosting model

Plot



Conclusion

First of all, by reading WHO latest weekly summary, senior people accompanying with own disease before this pandemic are especially vulerable to COV-19. Therefore, we sub-grouped the population by 5 age range, roughly covering from infants to teenagers, and elderly people. Then, we use Kmeans and spectral clusterting method to cluster the cov-19 data, K-means shows cov-19 case is positive correlated with the regions where economy is more developed, but in terms of death number, it does not show any particular pattern, it is evenly distributed around the country.

For classfication, SVM and Penalized logistic regression are used to study the area where the death number per 100,000 people greater than 1. The model can identify the hot spot area such as east coast New York and Chicago. But the less developed area, like west of US and mid of US did not observe a significant death.

In regression section, four different models are fitted and used to predict one week death number of 4/22 to 4/29. From the importance factor rank figures, it is obvious that the senior people especially for those older than 65 are most vulnerable to the COV-19 disease.Because they might have a lot of disease before this pandemic, such as diabetes, heart disease or stroke. This disease will greatly increase the risk of death once they are infected with the COV-19 virus. Smokers are not sensitive to this COV-19 disease compared to other groups. But this point needs further investigation, since this is a small dataset and the pandemic is also developing, more and more clinical study are going on to check the potential risk factor of affecting COV-19. On one hand, the population density is also positive correlated with death toll. Considering this COV-19 is proved to strongly transmitted between people, so urban cities are expected to have more cases, such as New York, and Chicago. This cities have a lot of enterntainment premises, when people accumulated in this closed atmosphere, the virus are likely to stay in the air longer and people inside it are more likely to get infected.

On the other hand, the proportions of local citizens who have been enrolled in Medicare is also a important player. Since, this represents how wealthy the local citizens are. Usually, the wealthier county have lower risk in death, since they have much better medical resources and cleaner living environment. Also, they have more hospitals and private doctors around their neighbourhoods. Last, the cov-19 cases one week before and death toll just one day before have a large effects in predicting the next day's death number. This is reasonable since this is a time series data, and the death toll are much auto correlated with case number. This point may be further studied by checking the auto correlated function and partial auto correlated function with certain time lags. Approximately, a one week lag from case number and death number are expected, since this cov-19 have one week to two weeks incubation, so that a majority of patients will appear with clear sympotoms and hospitalized into ICU one week after they have been diagnosed.

Finally, we have fitted the 92 days train data from Janurary 22 to April 22 which includes death and case information combined with health-related data and demographics data. All of the four models predict very well in train samples. Besides that, the extraopolation one week after 4/22 is given in previous section. We can find that there are three different patterns of growth rate of cov-19, resepetively exponential growth, sub-exponential growth and linear growth. These findings support our clustering results well.

Finally, a lot of precautions can be done to reduce the mortality and infection case number. Wash your hand frequently and wear a mask to prevent virus aerosols stayed in the air flow into your nose and mouth when you are breathing outside. Also, doing some exercise at home to increase your ability defeat the virus. Do not expose to the outside, and stay at home will also reduce your risk of infection. Cancel the parties if it is not necessary. In the end, hope everyone stay safe during this unusual pandemic.