County-level COVID-19 Analysis

Linglong Ma (lm15)

Tianqi Qu (tianqiq2)

Hui Yang (huiy4) - Leader

05/15/2020

County-level COVID-19 Analysis

Linglong Ma, Tianqi Qu, Hui Yang

Introduction

The COVID-19 pandemic has had far-reaching consequences beyond the spread of the disease itself and has left a huge social implications. While people have done lots of researches and had some knowledge regarding this pandemic, we always hope to devote more efforts to understand and quarantine it. With a general hope to have a better understanding of COVID-19 pandemic, we apply supervised and unsupervised learning methods to county-level data and try to figure out underlying pattern of the infection and make some prediction. In detail, our goal is to find out the vulnerable population, propose prevention measures to reduce mortality, and predict the death counts for each county. In this project, we develop four clustering methods to the demographics and health-related information and preform mean analysis based on different clusters. We define a new class variable to represent the levels of mortality (deaths per 100,000 population), and apply four classification methods to model the variable. We also use four regression methods to predict the number of deaths one week from Apr 22. Specifically, we take the number of deaths in neighboring county into consideration, and it has positive impact on performance of our models. The clustering and the classification results indicate that the population density is significant factor to the confirmed case and death counts and the population with other disease such as diabetes, heart disease, stroke, and smoking habit is more likely to be infected. The regression results provide us with a tool to predict short-term death counts, but the model is not suitable for long-term prediction. The hope is that if including some predictors regarding policies implications and perform modeling to each county respectively, we may get a better prediction. All the effort in the project may lead to a deeper understanding to the COVID-19 pandemic.

Literature Review

There have been many papers related to COVID-19 and its prediction. Yu's paper builds 5 predictors to forecast county-level death counts in the United States [1]. A separate-county exponential predictor uses only data from that county to predict deaths in that county. A shared-county exponential predictor uses data from all counties to predict death counts for individual counties. An expanded shared-county exponential predictor uses all data including cases and deaths in neighbor counties to do prediction. A demographic shared-county exponential predictor also includes county demographic and health-related predictive features. And finally, a separate-county linear predictor which uses only data from that county to build a linear model to predict the deaths in that county. In all, the Combined Linear and Exponential Predictors (CLEP), which combines county-specific exponential and linear predictors, has the highest accuracy in predicting most of the cases in the United States in the short term. Yuan's paper aims at predicting daily incidence and deaths of COVID-19 in the United States, and it uses the related terms to perform Pearson correlation test and general linear model to examine correlations and predict future trends, respectively [2]. Hao's paper makes prediction with the model based on the Eyring's Rate Process Theory and Free Volume Concept [3]. While all these methods can perfectly predict the future trend of COVID-19 to some extent, there still exist many uncertainties such as the highly dynamic nature of COVID-19 and randomness of human interactions that may affect the accuracy of prediction. For instance, Neil's paper puts that non-pharmaceutical interventions

[1] Curating a COVID-19 data repository and forecasting county-level death counts in the United States

[2] Trends and prediction in daily incidence and deaths of COVID-19 in the United States: a search-interest based model

[3] Prediction of Coronavirus Disease (covid-19) Evolution in USA with the Model Based on the Eyring Rate Process Theory and Free Volume Concept

Table 1: Hierarchical Clusters Distribution

Cluster 1	Cluster 2	Cluster 3	Cluster 4
4	3135	1	1

Table 2: Hierarchical Clusters Distribution after PCA

Cluster 1	Cluster 2	Cluster 3	Cluster 4
3136	1	1	3

(NPIs) such as staying at home and keeping social distance might have some impacts on the reduction of COVID-19 mortality [4].

Data

Description & Cleaning

The dataset used for this analysis is a subset of the COVID-19 data repository from Prof. Bin Yu's group, which contains 276 COVID-19 related variables for 3142 county-level observations. However, not all the variables are needed in our analysis, and the COVID19 dataset has 240 variables after deleting unnecessary columns. We found out that there are 7 variables that have missing values, in order to fix that, we calculated the Euclidean distance between each county using their latitude and longitude and then created a list called neighbor_county3 to display the nearest 2 counties of each county. For the two variables indicating the number of people enrolled in Medicare, we refilled them using the population of this county times the average percentage of Medicare enrollment in the two neighboring counties, if one of the neighbors has zero enrollment, then it will only be the population times the non-zero percentage. Similarly, we replaced the missing value of disease related variables and the ratio of votes in the presidential election with the average value of nearest two neighbor counties. Then we checked the dataset again, there is not any missing value now.

Analysis

Clustering

Data

Before clustering, we need to select some useful variables including demographics, health-related information and COVID-19 case/death counts. For the convenience of comparison between different counties, we divide some demographics variables by the total population to get their percentage, then divide health-related variables by total population divided by one hundred thousand to reduce the impact of different population bases. The range of total population is very large, meaning some counties have extremely small population and some have extremely large population. So, it is necessary to eliminate the influence of total population on other variables. And we also scale the data.

Hierarchical Clustering

We use demographics, health related variables and counts which are 14 variables in total to make hierarchical clustering with complete linkage. The result of dendrogram (see details in code) is not very good. It seems that there are some extreme values. If we select 4 clusters, over 3000 observations fall into cluster 2. Other clusters only have few observations. Thus, we need to make some improvement.

[4] Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand

Table 3: K-means Clusters Distribution

Cluster 1	Cluster 2	Cluster 3	Cluster 4
45	895	1071	1130

Table 4: Mean Analysis for K-means Clustering

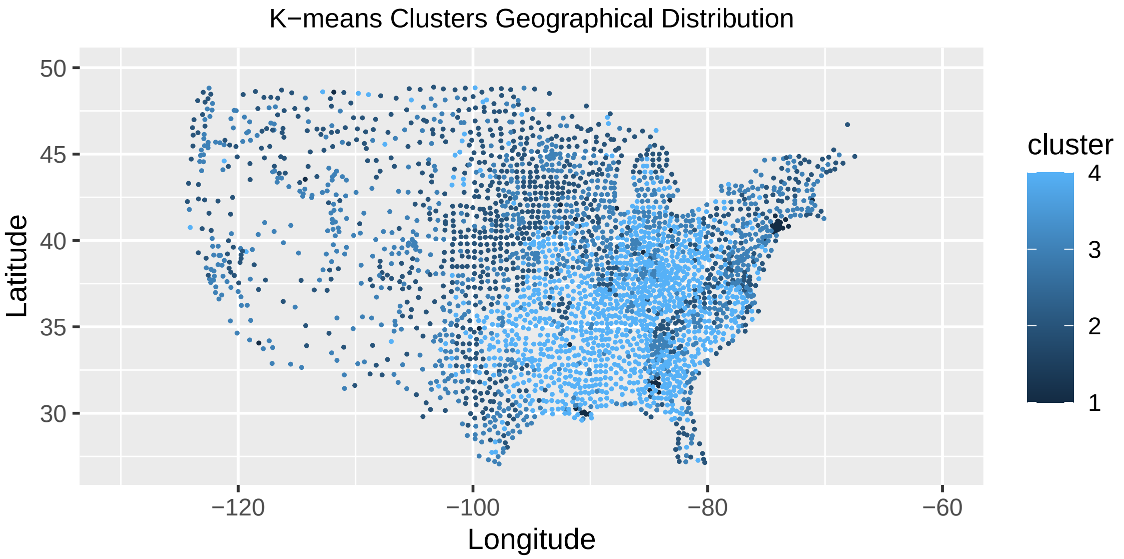
Cluster	Population	Density	> 65	Diabetes	Smokers	Heart	Stroke	ICU	Cases	Deaths
1	886782.98	5639.187	0.16	9.80	16.53	182.32	35.19	16.80	1577.12	76.73
2	30711.11	44.406	0.24	8.49	15.61	165.41	36.42	5.59	42.89	1.63
3	199920.19	377.364	0.15	8.82	15.71	161.86	36.69	17.64	107.78	3.90
4	39113.83	103.092	0.18	13.36	20.63	225.11	46.71	13.01	93.81	4.00

Consider many variables used in clustering have high correlations which may influence the result, we use PCA method to transform these variables to be unrelated and reduce the dimension. The dendrogram does not show much improvement after using PCA, and most observations still fall into the same cluster. Using other linkages also does not get a good cluster result. It seems that hierarchical clustering is not a good choice for our data.

K-Means Clustering

We choose 4 initial centers for K-means clustering. There are about 45 observations in cluster 1 which has the largest population density. It is reasonable because these big counties are just a small part. And other clusters have similar number of counties.

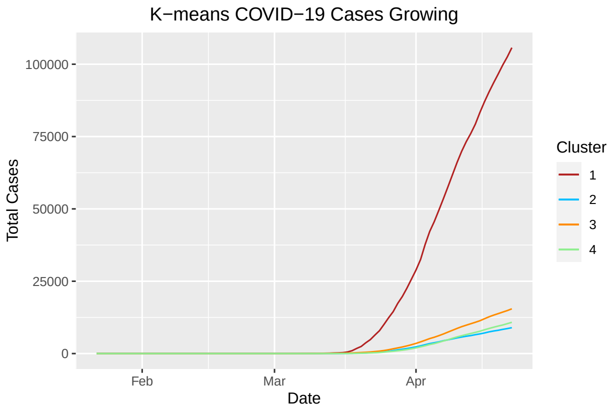
And we make a mean analysis based on different clusters. We can find that cluster 1 has the largest population density which leads to the most cases and deaths. Cluster 2 has the smallest population density and its cases and deaths are the least. Even though cluster 2 has more old people over 65 and less ICU beds, its COVID-19 counts are still not higher than others, meaning these factors do not have much influence on the total COVID-19 counts. Comparing cluster 3 and 4, we can find that cluster 3 has more population than 4 such that it has more cases. But the death counts of cluster 3 are a little smaller than 4. It may because people of cluster 4 have a higher proportion of diseases. So, these diseases can lead to a higher mortality rate.

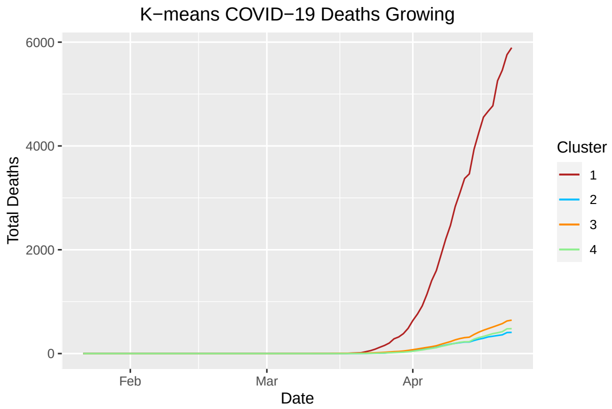


We can get the distribution of these clusters according to the longitude and latitude of counties. The figure shows that most high population density counties in cluster 1 are at the east and south of United States. Cluster 4 is at the middle area. Cluster 2 and 3 are at the west area.

Table 5: SOM Clusters Distribution

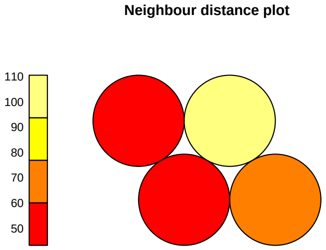
Cluster 1	Cluster 2	Cluster 3	Cluster 4
1049	965	1086	41

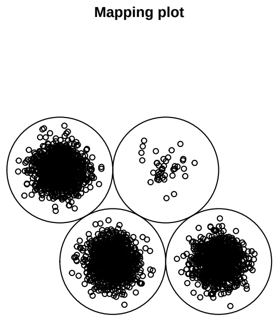


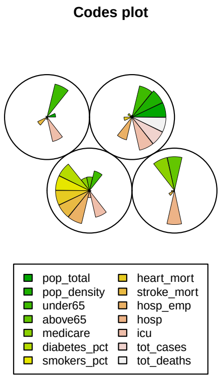


Then, we try to find the pattern of COVID-19 counts growing based on clusters. The growing rates of cases and deaths are decreasing from cluster 1, cluster 3, cluster 4 and cluster 2. It matches our previous analysis. The population of these four clusters is decreasing in the same order. We consider that the population is the most important factor for the COVID-19 cases and deaths growing.

Self-Organizing Map





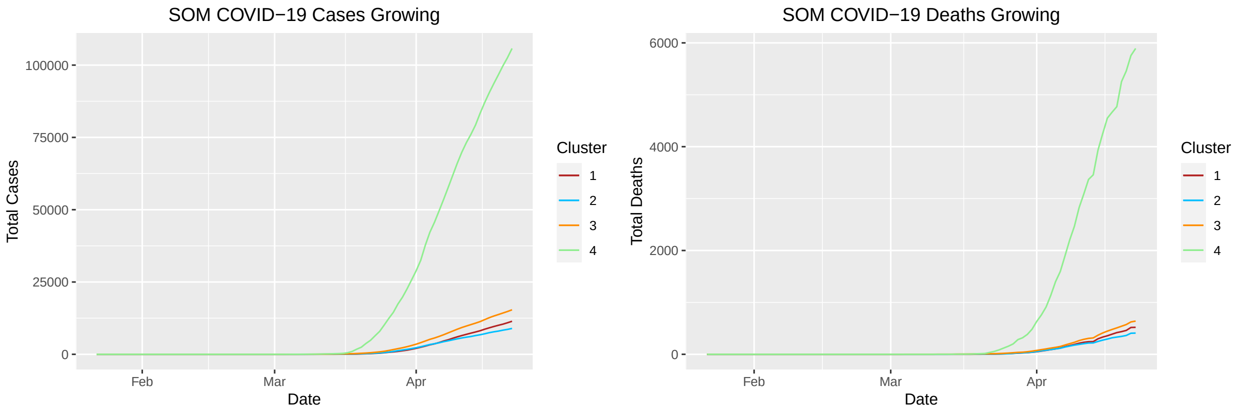


To be consistent with the previous method, we select the 2 × 2 grid for SOM. The number of observation of each cluster is also similar to K-means. The neighbour distance plot shows that cluster 1 and cluster 4 have higher neighbour distance indicating dissimilar. And low neighbour distance of cluster 2 and 3 indicates they are similar. Mapping plot shows how many objects are mapped into each cluster. The cluster 4 has the least number of objects and other clusters have similar number of objects.

Combining the codes plot and mean analysis, we can find that the result of SOM is similar to K-means. Cluster 4 has the largest population density which cause more cases and deaths. Cluster 3 has second largest population density so its cases are the second. But its deaths is less than cluster 1 because cluster 1 has more people with diseases. Cluster 2 has the least population so its cases and deaths are the least. More old people and less ICU beds do not lead to much more cases and deaths of cluster 2.

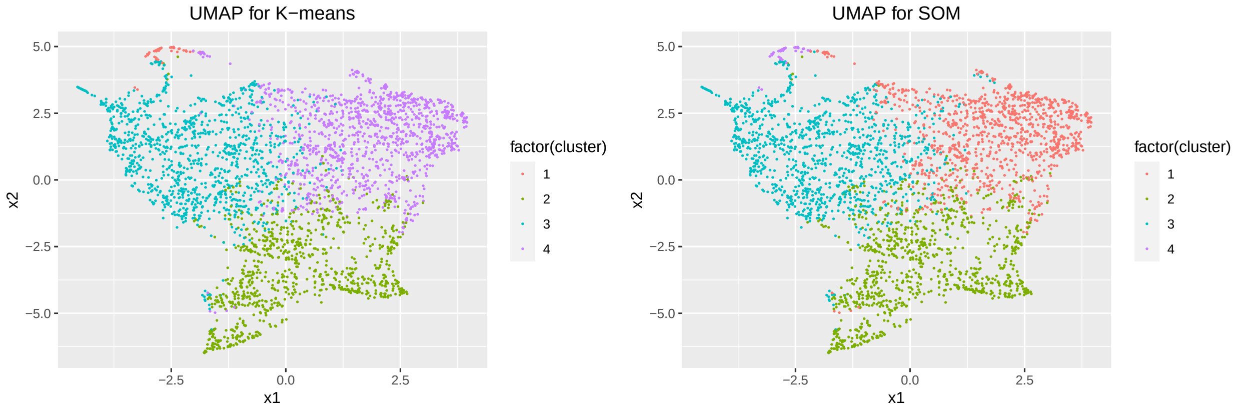
Table 6: Mean Analysis for SOM

Cluster	Population	Density	> 65	Diabetes	Smokers	Heart	Stroke	ICU	Cases	Deaths
1	38808.53	108.68	0.18	13.52	20.70	227.67	47.19	15.14	100.41	4.37
2	31012.46	49.81	0.24	8.60	15.75	166.70	36.48	6.01	43.58	1.64
3	198254.52	367.14	0.15	8.95	15.90	162.84	36.96	15.59	106.79	3.89
4	969835.39	6179.80	0.16	9.31	16.33	181.23	34.39	18.07	1644.34	78.09



As in the K-means analysis, cluster 4 which has the largest population has the highest cases growing rate, and the growing rate become smaller when the population is decreasing. The growing rate of cluster 4 is significantly higher than others, which can also prove the influence of large population on the growing of COVID-19. More people will cause more counts and higher growing rate.

Uniform Manifold Approximation and Projection



We can use UMAP to reduce the dimension of high-dimensional data and visualize it. By using UMAP method, we can project the points of high-dimensional space into low-dimensional space and retain the data structure as much as possible, then show the distribution of data in the low-dimensional space. After using UMAP, we can see the plot of K-means and SOM. The clusters distribution of these two methods are very close, which can also reflect the similar results of them.

Discussion

Table 7: Table: Accuracy for binary classification models

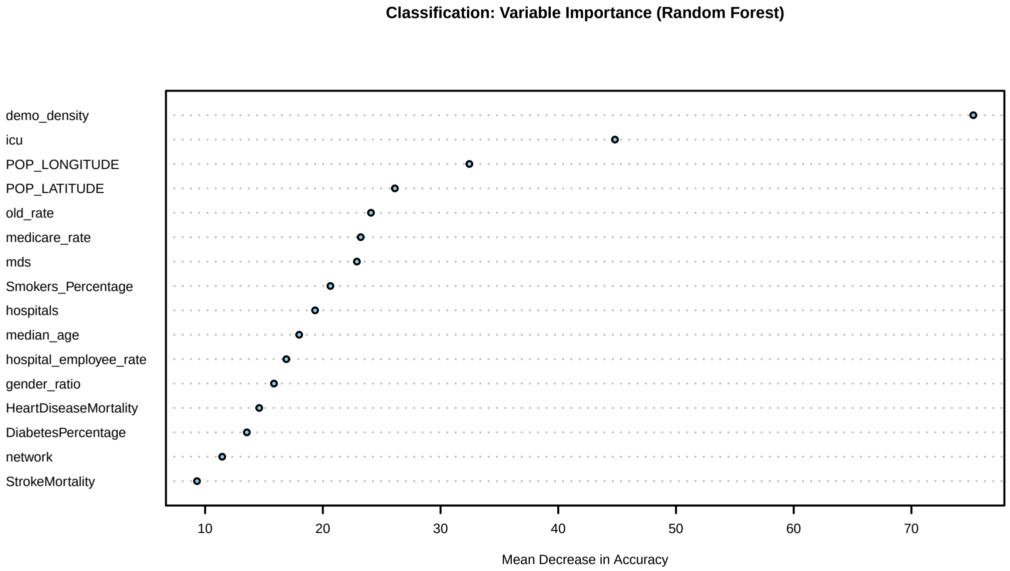
Model	Tuning	Training Accuracy	Testing Accuracy
KNN	k = 7	0.670	0.639
RDA	gamma = 0, lambda = 1	0.763	0.743
Random Forest	mtry = 9	0.765	0.743
Gradient Boosting	nrounds = 100, eta = 0.3	0.727	0.727

For total cases and deaths, more base population will lead to more cases and deaths counts. Besides, the health condition such as diabetes, smoke and heart diseases will also cause more death counts of COVID-19. For the growing of cases and deaths, the base population is also the most important factor. If there are more population or density, the growing rate will be far greater than other counties which have less population. It is reasonable because COVID-19 can be spread by contact. If there are more people in a county, the chance of contact between people will be more, so the risk of infection will be increased. And because of the exponential growth, the impact of large population will be more significant.

Classification

For this part, we develope a new variable to see whether the death per 100,000 people is greater than 1. The variables we selecte to do classification are geographical identifiers, demographics, and health-related predictors. We perform some modifications to each variable. For instance, in order to make them comparable between each county, we divide each variable by the total population of that county divides 100,000. For the number of hospitals and ICU beds, we simply scale the data. Then we separate the data into train and test datasets, the training set contains 80% of the whole dataset.

In order to model the new outcome, four different classification models were trained, which are k-nearest neighbors model, regularized discriminant analysis, random forest model and gradient boosting model, each using 10-fold cross-validation. The best tuning parameters were chosen using Accuracy.



For K-nearest neighbors, the tuning parameter is k, which represents the number of neighbors. Through 10-fold cross validation, our best tuning k is 7, and the test accuracy using the best KNN model is 63.9%. This method has the lowest testing accuracy, KNN assigns object to the class most common among its k nearest neighbors, it is a type of instance-based learning where the function is only approximated locally and all computation is deferred until function evaluation. This method is easy to understand and implement. However, A peculiarity of the KNN algorithm is that it is sensitive to the local structure of the data. As has been proved in cluster analysis, the number of counties in each cluster varies a lot, moreover, we cannot simply use mathematical distance to do classification, we real world situation is much more complicate, thus, KNN may not be a good method.

For Regularized Discriminant Analysis, the best tuning parameters are gamma=0 and lambda=1, which indicates that this is a linear discriminant analysis using a covariance matrix in the model. The test accuracy of this method is 75.6%. Linear discriminant analysis attempt to express one dependent variable as a linear combination of other features or measurements and then do the classification job. LDA works when the measurements made on independent variables for each observation are continuous quantities. This method has a particularly obvious advantage in dealing with the unbalanced pattern category, thus it has the highest accuracy.

In random forest method, the best tuning is 9 randomly selected predictors, the test accuracy is 73.5%, a bit lower than that of RDA. Random forest constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes or mean prediction of the individual trees. It can deal with huge number of variables and observations with high accuracy, it can also evaluate the importance of each variables. In this case, the most important variables selected make great sense. As we know, COVID-19 is a highly infectious disease, and counties with large population density are more likely to have more cases and deaths.The number of doctors and ICU beds indicates the overall medical level of this county, higher medical level will lead to less death.

The best tunings for extreme gradient boosting are 100 iterations and shrinkage of 0.3, the test accuracy is 72.7%. Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. The test accuracy is a bit lower than random forest and LDA, but still pretty good.

In conclusion, combining the test accuracy and property of these four methods, LDA has the highest classification accuracy, random forest has more real world meaning and can be understood better. They are all good classification methods.

Regression

Overviews

In this part, our goal is to figure out an approach to predict the number of deaths one week from Apr 22. The basic idea here is to fit several different statistical methods with both the demographics and health-related information, and the time series of COVID-19 deaths and cases counts, and choose the best one. Both linear and non-linear models are applied, and we can find some similarity between their results, which provide us with some enlightening views.

Here, we also acquire updated information at April 29 and use Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to evaluate our models.

Predictors

Since the raw data contains lots of redundant information, we start with variables selection. Based on above analysis and the paper by Prof. Bin Yu's group [5], four categories of variables regarding the demographics and health-related information are chosen, and we perform some extra processes to some of them to improve the comparability:

[5] Curating a COVID-19 data repository and forecasting county-level death counts in the United States

- County location and density: population density per square mile (2010), latitude corresponding to county's population center, longitude corresponding to county's population center
- County healthcare resources: number of hospitals (scaled) (2018-2019), number of ICU beds (scaled) (2018-2019)
- County health demographics: median age (2010), number of people eligible for Medicare in county per 100,000 (2018), percentage of the population who are smokers (2017), percentage of the population with diabetes (2016), deaths due to respiratory diseases per 100,000 (2017), deaths due to heart diseases per 100,000 (2014-2016)
- COVID-19 death counts: recent 5 days death counts, recent 1-day death counts in neighboring county.

To be more specific, we accept 8 out of 9 predictors from Section 3.4 of the paper, and scale the number of hospitals and number of ICU beds to suppress the difference between counties. We add latitude and longitude to represent the location of counties. Since the latitude and longitude are recorded corresponding to county's population center, it benefits to epidemic diseases models. In addition, we take the number of people eligible for Medicare in county into consideration because of the importance of the Medicare in the United States.

For the death counts, we only use the most recent 5 days of data. Since we cannot use ARIMA, we regard them as common variables and we update the data (replace the old data with the new one) everytime when we do the prediction. For example, we use the data from April 18 to April 22 to predict the number of death in April 23, and then we just accept the prediction and use the data from April 19 to April 22 together with this predicted death (April 23) to predict the number of death in April 24. The number of death in neighboring county is considered because of the infectiousness and migration. Here, we do not exactly define neighboring county geographically. In detail, since the policy vary from state to state, we only consider the neighboring county within state. The measure is the distance between counties' population centers, which can be calculated with the variables `POP_LATITUDE` and `POP_LONGITUDE`.

Out of convenience, we exclude the number of cases in this problem. On the one hand, we get almost the same accuracy when include them into our models. On the other hand, if we include those variable, we would have to construct another to predict the number of cases. In detail, when we predict the number of deaths from April 23 to April 29 in such case, we also need the value of the number of cases from April 23 to April 29, which we cannot get before April 22. Therefore, we can just ignore these variables.

For the model pattern, to guarantee that all the death counts are positive, we preform a log-transformation to deaths related predictors. The basic model pattern (linear) is shown as following:

$$log(Deaths_{t_0}) = \beta_0 + \beta_1 log(Deaths_{t_1} + 1) + ... + \beta_5 log(Deaths_{t_5} + 1) + \beta_6 P_1 + ... + \beta_{5+i} P_i,$$

where P_i are the set of demographic and healthcare-related features (such as population density, median age), $Deaths_{t_i}$ are the number of recent 5 days deaths.

Modeling & Discussion

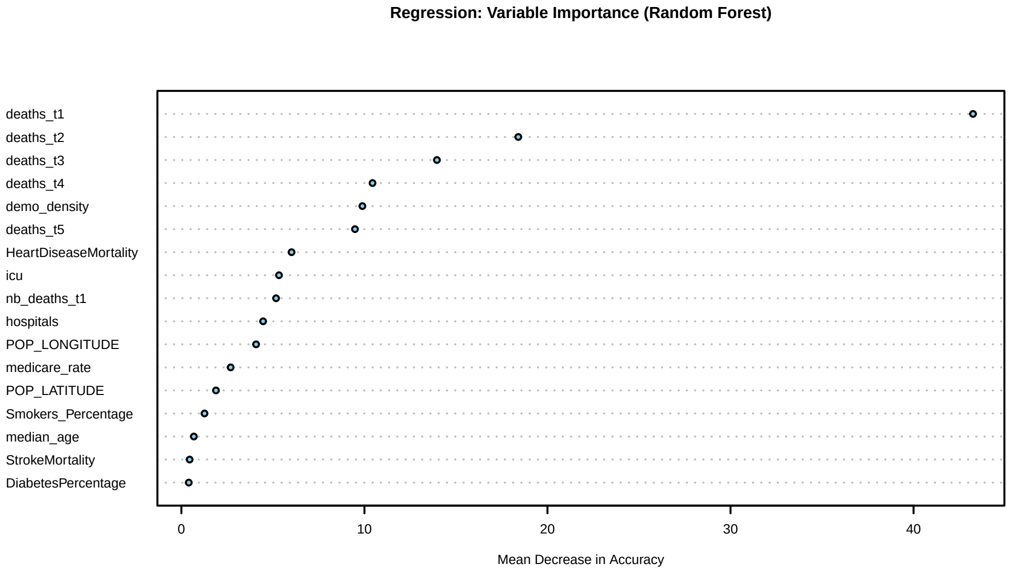
In order to predict the number of deaths, four different regression models are trained: random forest model, linear models, generalized additive model and k-nearest neighbors model. Out of bag samples and 10-fold cross-validation are used. The best tuning parameters are chosen using RMSE.

The k-nearest neighbors model may have bad performance because of the dimensionality issue, and we just consider it for comparison.

For the random forest model, we fit the training data via the `randomForest` package. The tuning parameter here is `mtry`, and out of bag samples are used to tune the best parameter. The result shows that the best `mtry` should be 9, which means that the best number of randomly selected predictors is 9. Random forest always perform well, but also the interpretation is usually more difficult. In order to illustrate, we can obtain the variable importance plot to see what set of predictors are important for us to do the prediction.

Table 8: Variable Importance (GLM | Non-zero part)

Predictor	deaths_t1	deaths_t2	deaths_t3
Overall	0.996018	4e-05	0.000252



The variable importance plot above represents the mean decrease in accuracy for each predictor. Fortunately, the results make sense intuitively. Initially, the number of recent 5 days deaths are of most importance. Since it is actually a time series problem, it is reasonable that the response is highly related to its lag orders. Apart from these predictors, population density, location (county's population center) and the number of recent 1 day deaths in neighboring county contribute more, and all the three predictors are related to close contact, which should be avoid during epidemic. The number of ICU beds represent the medical level of the county, which shares almost the same importance as the number of deaths in neighboring county.

For linear model, we fit the training data via the `glmnet` package, using 10-fold cross-validation to tune elasticnet mixing percentage `alpha` and regularization parameter `lambda`. If $alpha = 0$, the model would be Ridge regression model; if $alpha = 1$, the model would be Lasso regression model; if between 0 and 1, it would be Elastic Net model. The result turns out that the best tuning parameters are $alpha = 1$ and $lambda = 0$. Check for the coefficients, we can find that only `deaths_t1` has non-zero coefficient, which is about 0.97. The variable importance table shows that only recent 3 days deaths counts are important. Since Lasso is a pretty strict model, those predictors with less influence tend to be removed.

For generalized additive model, we fit the training data via the `mgcv` package, using 10-fold cross-validation to tune feature selection `select` and method `method`. The best parameter is that `select` = FALSE, and `method` = GCV.Cp, which means that each predictor can not be penalized to zero and using generalized cross validation (GCV) for unknown scale parameter and Mallows' Cp for known scale. Checking for variable importance table, still we get really high value in `deaths_t1`. Age issue is also important in the method.

For KNN model, the best tuning parameter is $k = 7$. Since the KNN model here is just for comparison, see detail interpretations in the `Prediction` part.

Table 9: Variable Importance (GAM | Greater than 1)

Predictor	deaths_t1	median_age	demo_density	POP_LONGITUDE
Overall	Inf	33.12	4.14	4.11

Table 10: Summary for Regression Models

Model	Random Forest	Linear (Lasso)	GAM	KNN
Tuning	mtry = 9	alpha = 1, lambda = 0	select = FALSE, method = GCV.Cp	k = 7

Prediction

With the updated data, we can check over the models by calculating testing RMSE and MAE. This updated dataset is also from Prof. Bin Yu's group have some counties missing in the dataset, but it is actually the best dataset we can obtain from Prof. Bin Yu's group.

As we have mentioned above, we need to update the data (replace the old data with the new one) everytime when doing the prediction. In detail, we renew the number of recent 5 days deaths and the number of recent 1 day deaths in neighboring county. Unsurprisingly, the bias of the predictions tend to become larger and larger, since we keep using inaccurate data.

Based on the results, we can find that the KNN model is not a proper method here. Both RMSE and MAE are much larger than other models. From the best tuning parameter k = 7, we know that KNN model use 7 nearest neighbor to predict the deaths number. However, such neighbor is the mathematical neighbors, not the geographical neighbors. Since our predictors contain both demographics & health-related information and time series, contain too much noise, we actually need to do some model selection, and KNN can not achieve this goal.

For the random forest model, linear model and generalized additive model, we get acceptable results. Overall speaking, generalized additive model have the best performance in both short-term (3 days) and long-term (7 days). Random forest model has relatively stable performance. linear model, Lasso, perform well in short-term, but become worse in long-term. According to above discussion, we know that only one predictor is left in the linear model, and that is why the long term predictions are bad here.

Although generalized additive model has better RMSE and MAE, we still want to choose random forest model as our final model. The first reason is that we need a stable model to help us do the prediction. The second reason if that random forest model can be explained and contains more real world meaning.

Table 11: Daily MAE for Regression Models

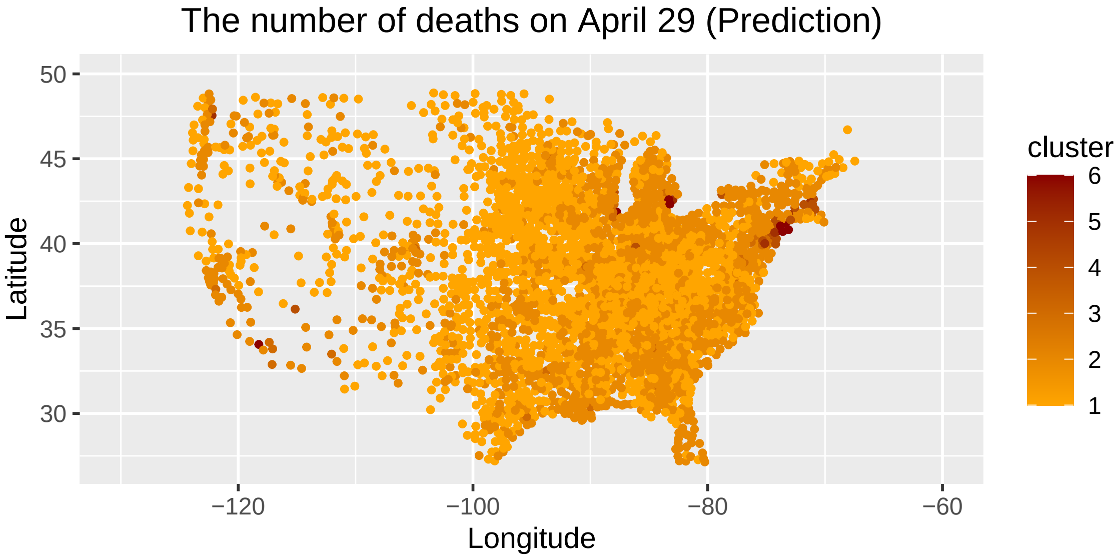
Model	April 23	April 24	April 25	April 26	April 27	April 28	April 29
Random Forest	1.321	1.187	1.507	1.850	2.054	2.338	2.723
Linear Model	1.474	1.576	2.262	2.919	3.324	3.860	4.638
GAM	1.145	1.208	1.552	1.908	2.039	2.267	2.659
KNN	9.957	9.957	10.494	10.984	11.270	11.650	12.278

Table 12: Daily RMSE for Regression Models

Model	April 23	April 24	April 25	April 26	April 27	April 28	April 29
Random Forest	14.363	14.671	16.531	18.999	20.459	21.839	24.088
Linear Model	10.952	11.836	15.851	20.083	22.863	25.860	30.367
GAM	8.467	14.601	16.768	19.181	20.374	21.693	23.962
KNN	67.538	67.538	70.742	74.100	75.911	77.925	81.555

Table 13: RMSE | MAE (7-Days) for Regression Models

Model	Random Forest	Linear Model	GAM	KNN
Total RMSE	130.952	137.810	125.046	515.309
Total MAE	12.980	20.052	12.777	76.589



Improvement

From the perspective of prediction, the are some methods to improve the accuracy. Initially, the model can be built for each county respectively. Differences in medical level, population and policies lead to gaps between counties, which undermine the effect of the data. For those unrelated counties, interaction from data produces negative side effect and make the predictions less accurate. To achieve the goal, we need more data for each county and we may consider to extend time periods to the beginning of the epidemic or the day with first death occurred.

Conclusion

From the result of previous analysis, we can find the most vulnerable population to corona virus and some feasible prevention measures to reduce mortality. All of previous analysis show that the population and diseases are important factors related to COVID-19 case and death counts.

According to the result of clustering, we consider people who live in the big cities where has more population density and people have diseases will be more vulnerable to COVID-19. It can make sense in the real-world situation. As we know, COVID-19 is a highly infectious disease. The higher population density a county has, the higher the incidence of disease among the population. The four diseases mentioned above all contribute to the risk of getting corona virus. The immunity and resistance of diabetic, heart disease and stroke patients are not as good as normal people. Besides, high blood sugar is easy to provide a very good environment for bacteria and germs, thus, people with diabetes are more likely to get infected. Smokers are likely to be more vulnerable to COVID-19 as the act of smoking means that fingers are in contact with lips, which increases the possibility of transmission of virus from hand to mouth. Smokers may also already have lung disease or reduced lung capacity which would greatly increase risk of serious illness. Conditions that increase oxygen needs or reduce the ability of the body to use it properly will put patients at higher risk of serious lung conditions such as pneumonia. Thus, people live cities with high density and have diseases will be more vulnerable.

For prevention measures to reduce mortality, we can also get some ideas from our analysis. Classification analysis shows some more factors related to COVID-19. One is the longitude corresponding to county's population center and the other is the number of medical doctors and ICU beds in each county. Most high population density counties such as New York are at the east of the United States. All these counties tend to have higher growth of the confirmed case and death counts. The number of medical doctors and ICU beds indicate the medical level of each county. Although there has not been an effective method to treat the corona virus, with good amount of knowledge and professional medical care, the death rate of this horrible disease may drop a little bit. For the regression part, we can see that the most significant variables are the number of death five days before. Same as the results of cluster and classification analysis, the population density of each county, the percentage of diseases, the number of hospitals and ICU beds can have a big influence.

Thus, we consider several ways to reduce mortality:

- Corona virus can spread rapidly among the crowd. Since the population in the county cannot be reduced, we can separate ourselves from others by avoiding going to crowded places such as restaurants and public parks and simply staying at home to prevent contact with others.
- Develop healthy living habits, through more exercise and healthier diets, we can reduce the risk of getting other diseases and enhance self-immunity at the same time.
- Medicare and the number of medical facilities are important to reduce deaths of corona virus. It is difficult to build more hospitals or create more ICU beds, however, many severe patients may need ventilator to assist breathing, thus, we can try to manufacture more ventilators or import from other countries to help reduce mortality.

There are still some deficiencies in our model and we can make some improvements in the future. We can build model for each county only using the data from this county and make a prediction for this county. Besides, we can consider more influential factors into our model. For example, the policy of stay at home can change the growth trend of COVID-19 counts. Because if people stay at home, it will reduce their contact with others such that reduce the risk of infection. Taking these factors into consideration will be helpful to improve the accuracy of our model.