The International Journal of Computational Science, Information Technology and Control Engineering (IJCSITCE) Vol.10, No.4, October 2023

MACHINE LEARNING CLASSIFIERS HELP TO MANAGE COVID-19 DISTRIBUTION IN CHINA

J. Wei[1],Y. Mingxuan[1], Z. Alsahfi[2], P. Norouzzadeh[3], T. Ye[2],
E. Snir[1] and B. Rahmani[3]

[1]Washington University, Olin Business School, St. Louis, MO, US
[2]Maryville University, Math and Computer Science Department, St. Louis, MO, US
[3] Saint Louis University, Computer Science Department, St. Louis, MO, US

ABSTRACT

The coronavirus disease 2019 (Covid-19) first appeared in Wuhan, China in December 2019. It spread very quickly from Hubei to the rest of China within only 30 days [1]. As of February 14, 2020, 78.91% of the total confirmed cases in China were in Hubei province which is in midland China and 60.33% of the total confirmed cases in Hubei were in Wuhan [2]. In this project we use K-Means clustering and regression analysis to classify the cities in China based on location and infection rate. The goal is analyzing the demographic and geographic characteristics of the clusters containing less than 5 members with high infection rate. We found that all the cities in the small groups with high infection rate are from Hubei province. We conclude that based on K-Means clustering and regression analysis, midland China is in trouble.

1. INTRODUCTION

Covid-19 is an infectious disease which is caused by the coronavirus. It is currently spreading worldwide at a faster rate than since it was first reported in Wuhan city in China in 2019 [3]. WHO declared Covid-19 as a world pandemic in March 2020 and as of now, there are almost 108 million confirmed cases and over 2.4 million deaths [4]. It causes a painful breathing disease, which has taken a large number of lives [3]. Cities in China have managed the spread of this virus [5]. This regional spread of the virus seems to have had a major effect on the social, political, and economic environment in different countries that have altered existing interactions [6].

To manage Covid-19 countries have enacted several measures such as travel bans, the use of masks in public places, and social distance [7]. In China, several measures were placed which include: active case surveillance, environment sanitation and disinfection, strict quarantine and follow-up, public awareness through reporting of confirmed cases [8]. There is also monitoring of people's body temperatures in the railway, airport, freeway toll station, subway entry, neighborhood entry point, and so on [8].

This research uses K-means clustering to define data subtypes in such a way that datasets are nearly identical in the same cluster [9].

Section 2 includes a literature review. We describe the data in section 3. Methodology, including regreesion analysis and k-means analyses, comes in section 4. We show the results and provide a discussion in section 5. In setion 6 we identify directions for future research. And finaly in section 7 we summarize the paper.

DOI: 10.5121/ijcsitce.2023.10402

2. LITERATURE REVIEW

Previous research into the intensity and spread of Covid-19 has employed various methodologies for modeling the disease. Early papers attempted to estimate the transition rates from infected individuals to others [10]. This early work focused on direct interactions to estimate possible transmission and exposure within a population. While these allows estimating how quickly Covid-19 may spread, it doesn't identify the actual risk to different localities, which is presented in this paper.

More recent papers focus on the spread of the virus, employing models based on the susceptible infectious-removed (SIR) model and extensions of these models [11], [12], [14]. These models explicitly track local mobility to assess the likelihood of transmission from one individual to another. Extensions to the model evaluate different public policy responses to infected individuals, [12], with a focus on hospital capacity and public awareness. The data required for these types of analyses is much more detailed than the proposed methodologies in this paper. Since detail mobility data from tracking individuals may not be viable, those methods are not always applicable.

Other papers investigate the spread of Covid-19 within China, similar to ours [13], [14]. Some of these focus on identifying the trajectory of spread from Wuhan to other cities in China early in 2020, before the disease was declared a pandemic [13]. Competing papers evaluate the impact of alternative lock-down policies in the spread of the virus. Neither of these classify cities based on their exposure risk, which is a contribution of this paper.

3. DATA DESCRIPTION

Covid-19 data is from all 329 cities in China from March 2020 to June 2020. The data was collected from WHO reports on the coronavirus and the National Health Commission (NHC) [4]. Features included are the population of each city in China, the city's location, and region. Total confirmed (T_c) cases data is included. The spread rate is calculated by dividing T_c by the city's population:

$$\frac{T_C}{population} \cdot 10{,}000 = Spread\ Rate,$$

4. METHODOLOGY

4.1. K-Means Clustering

We build a K-Means clustering model over 329 cities in China. The analysis divides cities into 20 clusters, based on the spread of Covid-19. We then compare these clusters to the geographic location of cities within clusters. The K-Means clustering algorithm divides the data into k-clusters such that similar cities are grouped together. Based on the measure of T_c we define the "distance" between cities. The algorithm minimizes distances within clusters, while maximizing the distance between clusters . Each cluster represents the mean value of cities within the cluster [15]. The cost function of the K-Means clustering is the distances between the observations in the cluster and centroid. The objective function finds the best k value that minimizes the cost function [16]. KMeans clustering method is applied to maximize the Silhouette values of cities and their respective cluster centers. The Silhouette is a method to evaluate the validity of clusters. It finds out the optimum k value by a ratio scale data (as the method of Euclidean distances does), which identifies well-separated cluster [17]. We measured the Silhouette values of all clusters with

different k values and chose the best k. Higher Silhouette values indicate better clustering. We found that when k equals 20, the Silhouette values are largest, indicating the best clustering within the Silhouette method.

K-Means is an unsupervised learning method. Labeling clusters is not important. In this study we used both supervised and unsupervised learning methods to make our results stronger.

4.2. Linear Regression

One approach to evaluating how geographic, demographic, and economic factors affect the number of cases of Covid-19 is through supervised learning using linear regression. We hypothesize that these three factors are related to the number of cases of the virus in by August 2020. The data for 329 cities in 8 geographic regions allows evaluating several such models.

In developing the regression model, we made two decisions. The first involves incorporating the geographic data. The second focuses on how to view the large heterogeneity in the number of cases of Covid-19. As seen everywhere in the world, Covid-19 appears to be highly concentrated in some cities, while other cities may exhibit only a few cases. One approach is to treat this variation as reasonable. The drawback of this approach is that cities with large concentration of Covid-19 have a substantial effect on the results of the model. Another view is that there are numerous outliers in the data. If the objective of the model is to evaluate average effects, outliers should be removed. Each approach has its merits. We present both sets of results in section 4.

5. RESULTS AND DISCUSSION

5.1. K-Means Clustering

Table 1 summarizes the 20 clusters of 329 cities. There are 11 clusters with less than 5 members.

Table 1: Information of 20 Clusters and 329 Cities

Cluster	Average Spread Rate	Number of Cities	Cluster	Average Spread Rate	Number of Cities
0	0.138998	57	10	0.077705	9
1	44.898323	1	11	2.073256	3
2	4.348363	2	12	7.148953	1
3	0.036970	52	13	3.106608	3
4	0.030999	49	14	1.720157	1
5	0.067883	55	15	5.884736	1
6	0.051303	35	16	0.378943	7
7	13.154666	1	17	0.912154	2
8	0.080822	34	18	0.650542	1
9	0.073778	14	19	0.710686	1

The bubble chart of Figure 1 shows the classification result. The diameter of the circle is proportional to the number of cities in the group. The numbers inside each circle is the spread rate value for the cities.

There are 6 clusters that have fewer than 5 cities and spread rate higher than 2.8. We would discuss the demographic and geographic characteristics of these clusters and explain why these cities have high spread rate. Below is the map of Hubei Province in Midland China (Figure 2), where most COVID-19 cases were confirmed, the map illustrates the geographic distribution of cities in Hubei Province [18].

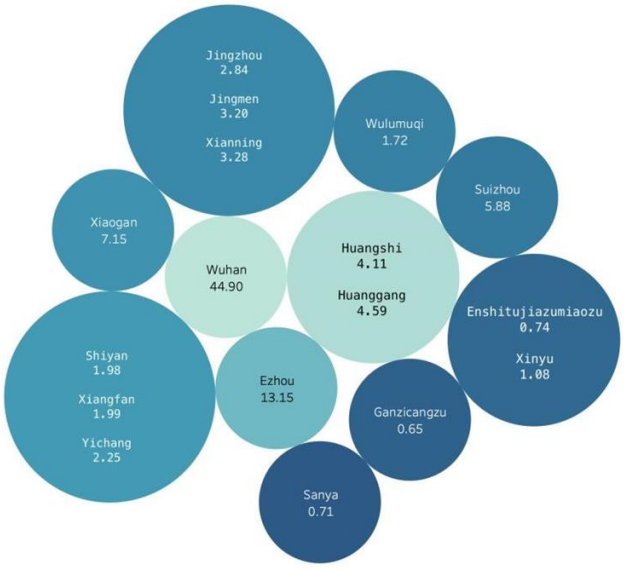

Fig.1: Bubble Chart of Small Clusters (Less Than 5 Members)

Table 2: Clusters with Fewer than 5 Cities with Spread Rate above 2.8 (per Thousand People)

City	Province	Location	Population (1,000)	Total Cases (August 2020)	Spread Rate	Tourist	Cluster
Wuhan	Hubei	Midland	11,212	50,340	44.89832	1	1
Huanggang	Hubei	Midland	6,333	2,907	4.590242	0	2
Huangshi	Hubei	Midland	2,471.7	1,015	4.106485	0	2
Ezhou	Hubei	Midland	1,059.7	1,394	13.15467	0	7
Xiaogan	Hubei	Midland	4,921	3,518	7.148953	0	12
Jingzhou	Hubei	Midland	5,570.1	1,580	2.836574	0	13
Jingmen	Hubei	Midland	2,897.5	928	3.202761	0	13
Xianning	Hubei	Midland	2,548.4	836	3.28049	0	13
Suizhou	Hubei	Midland	2,221	1,307	5.884737	0	15

Fig.2: The Map of Hubei Province in Midland China

Cluster 1: Wuhan

Wuhan is the only city in Cluster 1. Wuhan is the first city in China where Covid-19 was identified. [19] Wuhan has the largest Spread Rate of around 45 (per ten thousand population). Wuhan is characterized by a large population of 11 million and high population density, Wuhan is exposed to the risk of rapid spread of Covid-19 cases. As a transportation center in Midland China, Wuhan has a well-connected network of transportation. People from nearby cities pour into Wuhan for employment and education opportunities. Another reason that causes the large number of cases is that Wuhan is a tourist city. Travelers increase the regional density of population, thus making it easy for Covid-19 to spread in Wuhan.

Cluster 2: Huangshi and Huanggang

These 2 cities are in the same province as Wuhan (Hubei Province) and both are in Midland China. Though the population of Huanggang is about three times that of Huangshi, by August 2, 2020, both cities have similar spread rates of 4.1-4.6. Neither of them is a tourist city, so the transient population of the two cities tend to be small. However, when we look at the geographic location within the Hubei Province of the two cities, we find that Huangshi and Huanggang lie at the two sides of the Yangtze river, which is the longest river in China.

Cluster 3 includes 52 cities with similar spread rates, tourism development, and geographic location.

Cluster 7: Ezhou

What distinguishes Ezhou from other clusters is its middle-sized population and high spread rate. This may be because Ezhou is geographically close to Wuhan, the capital city of Hubei Province, and thus many people go to work in Wuhan during the day and return to Ezhou where they live.

As a result, the city has very large number of confirmed cases of Covid-19.

Cluster 12: Xiaogan

As a city of Hubei Province, Xiaogan lies on the northern part of Yangtze River. Xiaogan and Wuhan are located across the Yangtze River. Xiaogan is a mid-sized city with a population of around 5 million, but by August 2020, 3,518 cases were confirmed. It has a large spread rate of 7.

Similar to Ezhou in cluster 7, Xiaogan is in Cluster 12 alone because of the change in the infected population in the short time span and the relatively large spread rate. Just as Ezhou, Xiaogan is grographically close to Wuhan.

Cluster 13: Jingmen, Jingzhou, Xianning

All three cities are in Hubei Province in Midland China. None of them is a tourist city. Jingmen and Xianning have similar populations, while Jingzhou has twice the population, at around 1 million people. The number of confirmed cases in Jingzhou is twice that of the other two cities, yielding a spread rate for the three cities of around 3.

Cluster 15: Suizhou

Suizhou is 200 kilometers away from Wuhan. [20] Compared with Xiaogan, Huanggang and Huangshi, Suizhou is far away from Wuhan. According to the statistical yearbook of 2019 published by Hubei Province, Suizhou City has a population of 2.22 million, ranking it 12th in the province. Suizhou's GDP is 10.11 million yuan, ranking 11th in the province[21]. The spread rate in Suizhou was substantial, even early in the pandemic. It's ranked fourth in Hubei Province.

The characteristics of cities with large number of confirmed cases in China are as follows: 1. Close to Wuhan; 2. Large population; 3. Frequent contact with Wuhan. Suizhou is a city in Hubei Province, which does not have all of the above three characteristics, but still has a large number of confirmed cases. This underscores the complexities of managing a virus such Covid-19, and the need for statistical modeling to identify drivers of the spread of the virus.

5.2. Linear Regression

The data includes 329 cities in 8 geographical regions. By looking at the distribution of cities to regions, we chose the Southwest region as the base category. One justification is that it has a fairly large proportion of the cities at 16%. It also has one of the lowest number of cases at 27.87 in an average city. Our linear regression model forecasts the number of cases as a function of population (in tens of thousands), a binary variable for whether the city enjoys strong tourism, and indicator variables for the remaining 7 regions.

We present 2 regression models. Model (a) includes all 329 cities. In Model (b) we use an iterative approach to removing outliers. Since the number of Covid-19 cases is highly concentrated in a few cities, these cities can be viewed as outliers. If the purpose of the analysis is to evaluate average effects of the number of cases of the virus, outliers should be removed. As seen in Table 3, the differences between the models are substantial.

The iterative method for removing outliers developed several regression models. In each step we identified those cities with Studentized Residuals more extreme than ± 3.5. These observations are deemed outliers. (This is justified under the assumption of a Normal distribution for the errors.) These outliers are removed and the regression is developed again. In the first step, only one observation is removed (Wuhan). In the second step, 5 more observations are removed. In all, 43 cities are removed in 14 iterations.

The two models present very different pictures regarding the number of Covid-19 cases in China. Both models conclude that Midland China experiences more Covid than the Southwest region.

They also both recognize that tourism is an important factor in the spread of the virus. That's about all the models have in common. Even the magnitude of the coefficients is substantially different in both models.

Based on Table 3, from Model (a) one concludes that the average number of Covid-19 cases in a city in the Midland region is 1,000 more than in the Southwest region. We reject the null hypothesis, although only at a p-value a bit below 0.05. Here the null hypothesis evaluates whether the number of cases is the Midland region is identical to the number of cases in the Southwest. The conclusion is that there is a statistically significant difference.

In contrast, Model (b) finds that the difference between the Midland region and the Southwest is only 23.8 cases, in an average city, but it significant even at a very low p-value. The difference is quite stark. The reason for these differences is that Model (b) does not include the cities with the highest number of Covid-19 cases, including Wuhan. Removing these outliers may give us a better identification for the average effects of different factors.

Table 3: Regression Results. Model (b) does not include outliers.

Dependent Variable: Covid-19 Cases		
Independent Variable	*Model (a) Coefficients*	*Model (b) Coefficients*
Intercept	-151.89515 (419.64372)	-6.66584 * (2.64409)
Population (Ten Thousand)	0.09701 (0.55463)	0.06062 *** (0.00463)
Tourist	2,527.18292 *** (732.29303)	18.48986 *** (5.17232)
East	-152.77913 (521.0449)	7.90774 * (3.30545)
Midland	1,000.67316 * (497.36799)	23.80795 *** (3.33324)
North	7.57021 (596.1755)	-2.03427 (3.60732)
Northeast	74.44107 (599.31177)	8.7699 * (3.57386)
Northwest	102.21318 (543.50472)	2.70799 (3.23583)
South	-364.36751 (804.53318)	4.85477 (4.77519)
Northeast	-138.83986 (982.29534)	10.73823 (5.80928)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (Standard Errors in parentheses)

Looking at regional differences, Model (b) also identifies that two more regions have statistically significantly more cases than the Southwest. These are the East, at almost 8 more cases in an average city, and the Northeast, and almost 9 more cases. It may be interesting to note that both models do not identify significant differences for other regions in China.

The other common significant factor in both models also emphasizes the differences in the models. Model (a) suggests that a tourist city has 2,527 more cases than one without tourism. In model (b) the result is substantially subdued at only 18.5 more cases. Again, removing Wuhan and a few large outliers is at the heart of these differences.

The one other finding from Model (b) is that population is a statistically significant driver for the number of cases. Larger cities do have more cases of Covid-19, at approximately 1 more case per 10,000 population. We now see that the result is significant at a very low p-value. This result isn't observed in Model (a), because Wuhan and other moderately-sized cities incur the vast majority of Covid-19 cases in China. For example, Wuhan has over 50,000 cases, while Beijing and Shanghai, much larger cities, have less than 2,500 cases each.

From a statistical perspective, Model (b) is much better, which shouldn't be a surprise. Model (b) has few outliers. For Model (a) the R^2 is 0.072, while the RMSE is 2,722.8. In Model (b) these are 0.584 and 15.92, respectively. Overall, both models are statistically significant.

6. FUTURE RESEARCH AND IMPLEMENTATION

As Covid-19 continues to be a global pandemic, understanding the prevalence of cases continuities to be a challenge. In some countries, like China, the reaction to increases in cases is lock-downs. Other countries still need to be concerned about the impact the pandemic may have on available hospital beds. This research identifies the heterogeneity in Covid-19 infections. Both models recognize that infections in China were highly concentrated.

K-Means models identify similarities and differences across regions. It is important to evaluate the model over multiple time periods. This allows measuring how similarities across regions vary over time. If the model returns similar results at different time periods, it suggests certain cities are likely to have recurring outbreaks. More likely, future iterations of the analysis will lead to different clusters across time. This suggests that real-time monitoring of clusters is essential to limiting Covid-19 spread. Moreover, through real-time monitoring, policymakers can identify both current and potential hotspots.

The linear model generates similar results. In addition, the effect of tourism on the spread of Covid19 is statistically significant. Supervised models facilitate generating inferences, which support public policy. Continuous reevaluation of the linear model would provide insight into the importance of outliers in the spread of the virus. Similar to the unsupervised model, identifying whether the spread of Covid-19 is likely to be localized, or similar across regions, should drive policy decisions.

7. CONCLUSION

The regression results confirm the need for Cluster Analysis. From the regression, we see that traditional supervised models identify many cities as outliers. The Cluster Analysis is an effective way to deal with outliers. The multiple clusters capture substantial differences across cities. Many of the outliers removed in Model (b) of the regression are clustered into a few different groups. For example, the extreme number of cases in Wuhan lead to it being in its own cluster. Other cities, near Wuhan, were also clustered together.

The regression model is effective at identifying average trends in the data. After removing outliers, we see the importance of city size on the spread of the virus. Tourism, which is

hypothesized to be important in the cluster analysis, is now quantified. Even after removing outliers, tourism cities have 18 more cases than cities without tourism. The regression model also accentuates regional differences. Some of the results from the regression model are not highlighted in the cluster analysis.

In summary, unsupervised and unsupervised learning methods complement each other when investigating the spread of Covid-19. With frequent "bursts" in the spread of the virus, unsupervised models provide insights into the outliers. Those cities with very high or very low spread are identified. Conversely, there is also a need to understand the average spread of Covid19. For that, a regression model is useful. Together, these types of models help policymakers effectively allocate limited resources to combat Covid-19.

DECLARATIONS

AUTHORS' CONTRIBUTIONS

Wei, Minxuan, Alsahfi: Modeling, coding, writing the paper
Norouzzadeh, Ye: review and added new section
Snir: review and co-adviser
Rahmani: review and co-advisor

AVAILABILITY OF DATA AND MATERIALS

Data are available to submit if it needed. The data was collected from WHO reports on the coronavirus and the National Health Commission (NHC)

REFERENCES

[1] The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. The Epidemiological Characteristics of an Outbreak of 2019 Novel Coronavirus Diseases (COVID-19) — China, 2020. China CDC Weekly (2020).
[2] Zhu ZB, Zhong CK, Zhang KX, Dong C, Peng H, Xu T, Wang AL, Guo ZR. Epidemic trend of COVID-19 in Chinese mainland. Chinese Journal of Preventive Medicine. (2020)
[3] A. G. Hadi, M. Kadhom, N. Hairunisa, E. Yousif, and S. A. Mohammed, "A review on COVID-19: Origin, spread, symptoms, treatment, and prevention," Biointerface Res. Appl. Chem., vol. 10, no. 6, pp. 7234–7242. (2020)
[4] CCDC, "Distribution of new coronavirus pneumonia," Oms, vol. 2019, no. February, pp. 1–7 (2020)
[5] M. Azarafza, M. Azarafza, and H. Akgün, "Clustering method for spread pattern analysis of corona-virus (COVID-19) infection in Iran," J. Appl. Sci. Eng. Technol. Educ., vol. 3, no. 1, pp. 1–6. (2020)
[6] M. Azarafza, M. Azarafza, and J. Tanha, "COVID-19 Infection forecasting based on Deep Learning in Iran," Medrxiv, pp. 3–9. (2020)
[7] O. H. World, "Covid-19 Strategy Up to Date," no. April, p. 18. (2020)
[8] T. L. Xu et al., "China's practice to prevent and control COVID-19 in the context of large population movement," Infect. Dis. Poverty, vol. 9, no. 1, pp. 1–14. (2020)
[9] V. Chandu, "Identification of spatial variations in COVID-19 epidemiological data using K- Means clustering algorithm: a global perspective," medRxiv, p. 2020.06.03.20121194. (2020)
[10] M. Chen, T. Rui, J. Wang, et al. "A mathematical model for simulating the phase-based transmissibility of a novel coronavirus," Infect. Dis. Poverty vol 9 no. 24 p. 2 – 8 (2020).
[11] M. Liu, R. Thomadsen, and S. Yao, "Forecasting the spread of COVID-19 under different reopening strategies," Sci. Rep., vol. 10, pp. 1-8 2020
[12] A.. Ajbar, R. Alqahtani and M. Boumaza M "Dynamics of an SIR-Based COVID-19 Model with linear incidence rate, nonlinear removal rate, and public awareness," Front. Phys., vol 9, no. 13, pp. 1-13 (2021)

[13] Z. Du, L. Wang, S. Cauchemez, et al. "Risk for transportation of coronavirus disease from Wuhan to other cities in China," Emerging Infectious Diseases, vol. 26, no. 5, pp. 1049- 1052 (2020)

[14] M. Li, G. Sun, J. Zhang, et al. "Analysis of COVID-19 transmission in Shanxi Province with discrete time imported cases," Mathematical Biosciences and Engineering, vol. 17, no. 4 pp. 3710-3720 (2020)

[15] Theodoridis, S., Pikrakis, A., Koutroumbas, K., Cavouras, D. An Introduction to Pattern Reccognition : A MATLAB Approach. Academic Press, USA. (2010)

[16] J.B. MacQuuen: Some methods for classification and analysis of multivariate observation, in: Proceedings of the 5th Berkley Symposium on Mathematical Statistics and Probability, pp. 281–297 (1967)

[17] Thinsungnoen, T., Kaoungku, N., Durongdumronchai, P., Kerdprasop, K., & Kerdprasop, N. The Clustering Validity with Silhouette and Sum of Squared Errors. The Proceedings of the 2nd International Conference on Industrial Application Engineering 2015. doi:10.12792/iciae2015.012 (2015)

[18] China Highlights. (2020, September 26). Hubei Map. Retrieved from China Highlights|Travel Guide: https://www.chinahighlights.com/hubei/map.htm

[19] MA. Shereen, S. K. COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. Journal of Advanced Research, 91-98. (2020)

[20] Yan. D, Wang. W, Fang. X: Magnetic levitation in Suizhou, Hubei Province. CHSSCD (2008)

[21] China Statistical Publishing House: Hubei Statistical Yearbook in 2019. (2020)