Relative Severity Analysis and Time-series Prediction of COVID-19 Outbreak

Tanran Zheng
Institute of Industrial Economics, Jinan University
Guangzhou, China
zhengtr@gmail.com

Abstract: COVID-19, a highly infectious disease caused by virus named SARS-CoV-2, is spreading globally. In order to better understand and possibly contain this ongoing pandemic, this paper conducts a series of data analysis of COVID-19 time-series data of countries and territories to measure and predict the severity of the outbreak. First of all, we develop an index, Relative Severity Score, which measures and quantifies the relative severity of the COVID-19 for each country since its outbreak. Then we conduct regression analysis with this index and other geographical data, which shows that the mean of population age, average humidity, average temperature and average wind speed are statistically significantly correlated to the Relative Severity Score. At last, by performing logistic regression analysis using the Relative Severity Score, we make a prediction to the future trends of the pandemic in the countries that are currently experiencing accelerated outbreaks.

Key Words: COVID-19; sequential clustering and dimension reduction algorithm; time-series analysis; regression analysis; supervised machine learning;

1 Introduction

Coronavirus pneumonia (COVID-19) has broken out in many countries around the world. On February 29, 2020, the epidemic was defined by the World Health Organization as a global pandemic. As of March 30, Covid-19 has caused more than 750,000 infections and 35,388 deaths worldwide, with 201 countries and regions having confirmed cases [1]. Since the outbreak of the epidemic, there have been literatures and studies to model and predict the spread and development of the epidemic. However, due to the estimation of model parameters involved in most infectious disease models, the prediction accuracy is limited. Therefore, this paper hopes to study an evaluation of the severity of the epidemic trend based on the real spread of the epidemic and predict the future trend.

Preliminary analysis showed that although there were time differences in outbreaks among countries and regions, the trend of the epidemic situation was highly similar. Therefore, this paper can quantify and compare the severity of the epidemic situation in different countries and regions after the outbreak, namely the Relative Severity Score of the epidemic, by putting the epidemic situation data of various countries and regions on the same time axis for cross-sectional analysis.

Index of Relatively Severity Score of the epidemic can compare and judge the development of epidemic situation in different countries and regions. Some studies have pointed out that the severity of the epidemic also

depends on objective factors in different regions [2]. Therefore, after obtaining the Relatively Severity Score of the epidemic, this paper further analyzed the influence of geographical, human and social factors (such as average temperature, population density, health and medical resources, etc.) on the severity of the outbreak. Through the Ordinary Least Squares regression analysis [3], we found and judged the factors that have significant influence on the spread or severity of the epidemic statistically.

Finally, according to the Relative Severity Score of the epidemic, this paper used Regression Analysis in supervised machine learning model to predict the future trend of the epidemic.

2 Method

2.1 Sample Selection

In the aspect of sample selection, in order to ensure the comparability between samples, this paper selected countries where COVID-19 has already started to spread or has imported cases as samples. According to the data of Johns Hopkins University on the number of confirmed cases as of March 31, 2020 [1], this paper selected countries with a cumulative number of confirmed cases exceeding 100 as data samples. When calculating the Relatively Severity Score of the epidemic situation, as the epidemic situation in Hubei Province of China and New York State of the United States in the sample is very serious compared with their respective countries, they are the key areas for the outbreak of the epidemic situation, and the cumulative number of confirmed cases has always accounted for more than 40% of the confirmed cases in their respective countries. Moreover, China and the United States have large land areas and relatively scattered epidemic areas, so this paper listed the data of these two areas separately for analysis.

2.2 Selection of Data Analysis Time Period

Due to the different time of concentrated outbreak in each country, in order to horizontally compare the outbreak situations, this paper uniformly regarded the confirmed number reaching 100 as the sign of concentrated outbreak of the epidemic, starting from this date, and analyzed the trend of the epidemic situation in the next 40 days. For example, the number of confirmed cases in Hubei Province of China broke 100 for the first time on January 19, reaching 121, so for Hubei Province, January 19 was day-1; The number of confirmed cases in Italy broke 100 for the first time on February 24, reaching 132, so February 24 was day-1 in Italy. Through this method, the time-series data of outbreaks in different periods can be aligned, and the severity of outbreaks can be compared and analyzed on the same cross section.

2.3 Constructing Indices to Quantify the Severity of Epidemic

In order to make a horizontal comparison of the epidemic situation in each country, it is necessary to select indices to quantify the severity of the epidemic. The number of confirmed cases, the number of deaths, the number of cured cases and the number of newly confirmed cases are the main data used to disclose the epidemic situation in the world. However, the cross-sectional comparability of these three data is not high, because the population base and the number of confirmed cases base of each country and region are quite different, and there is no information directly covering the outbreak speed of the epidemic. Therefore, this paper chose to construct 11 derivative indices

with horizontal comparability according to these three data, and used these 11 derivative indices to make horizontal comparison on the epidemic area.

Table 1. Factors of severity of the COVID-19[1]

Indices	Abbreviation	Description	Calculation method
Confirmed Per Million	confPC	Cumulative number of confirmed cases per million people	Cumulative number of confirmed cases/total population
Confirmed DoD	confDD	Growth Rate of Daily Confirmed Number	(Cumulative number of confirmed cases/Cumulative number of confirmed cases on the previous day)-1
Death DoD	deathDD	Daily death toll growth rate	(Cumulative Deaths/Cumulative Deaths on Previous Day)-1
Recovery DoD	recDD	Growth Rate of Daily Cure Number	(Cumulative number of people cured/Cumulative number of people cured on the previous day)-1
Fatality Rate	fatalRate	Fatality rate	Cumulative Deaths/Cumulative Confirmations
Recovery Rate	recRate	Cure rate	Cumulative number of cured/cumulative number of confirmed cases
Remain Per Million	remPC	Existing infections per million people	(Cumulative number of confirmed cases-Cumulative number of deaths-Cumulative number of cured cases)/total population
Remain Ratio	remainRatio	Ratio of existing infections to cumulative confirmed infections	(Cumulative number of confirmed cases-Cumulative number of deaths-Cumulative number of cured cases)/cumulative number of confirmed cases
New Confirmed DoD	newConfDD	The growth rate of the number of newly confirmed cases on that day	(Number of newly confirmed cases on the same day/Number of newly confirmed cases on the previous day)-1
New Death DoD	newDeathDD	The increase in the number of new deaths on that day	(Number of new deaths on the same day/Number of new deaths on the previous day)-1
New Recovery DoD	newRecDD	The growth rate of the number of newly cured people on that day	(Number of newly cured persons on the same day/Number of newly cured persons on the previous day)-1

2.4 Model method

2.4.1 Sequential Clustering and Dimension Reduction Algorithm

First of all, in order to ensure the neatness of the source data, the model carries out data preprocessing on all derived indices introduced in Section 2.3, namely Outlier filling, Median Absolute Deviation (MAD) [4] and Z-Score standardization, ensuring that all indices are free from the interference of outliers, and converting all indices into the same dimension [5]. Due to the large number of indices, and the 11 indices are basically derived from the original data of confirmed number, death number and cure number, there is also a certain multicollinearity among the indices. Therefore, after preprocessing the data, the model uses Sequential Clustering and Dimension Reduction Algorithm [6] to cluster and reduce the dimensions of the indices to eliminate multicollinearity among the indices, so that the subsequent regression model and prediction model can better extract the effective information of the indices [7].

Sequential clustering and dimensionality reduction algorithm is an algorithm for processing multi-dimensional time series. The algorithm mainly uses the ideas of K-nearest neighbors algorithm and greedy algorithm for reference to cluster time-series data step by step in a certain order. After obtaining the clustered grouped data, step-by-step Ordinary Least Square regression analysis is carried out in each group, and the obtained residual error is taken as new data. Finally, the data in each group are weighted and averaged to form a new group of data in each group, and finally the dimension reduction effect is achieved [6].

[1] Source: Johns Hopkins University CSSE, World Bank

Firstly, the correlation of the original indices was analyzed. As can be seen from Table (2), the time-series correlation of some indices was greater than 0.6, and there was high multicollinearity. For example, the correlation between the growth rate of the number of confirmed cases on that day and the growth rate of the number of deaths on that day was as high as 0.92, which reflected the speed of the outbreak of the epidemic. The correlation between the cumulative number of confirmed cases per million people and the number of existing infections per million people was also as high as 0.94 during the concentrated outbreak. So, multicollinearity between these factors needed to be dealt with.

Table 2. Correlation matrix of original data before implementing sequential clustering and dimension reduction algorithm

Original index	confPC	confDD	deathDD	recDD	fatalRate	recRate	remPC	remainRatio	newConfDD	newDeathDD	newRecDD
confPC	1.00	0.56	0.54	-0.33	0.20	0.38	0.94	0.29	0.11	0.35	-0.29
confDD	0.56	1.00	0.92	-0.27	0.16	0.41	0.59	0.20	0.11	0.55	-0.45
deathDD	0.54	0.92	1.00	-0.32	0.05	0.46	0.56	0.20	-0.04	0.66	-0.34
recDD	-0.33	-0.27	-0.32	1.00	0.29	-0.21	-0.35	-0.46	-0.20	-0.37	0.39
fatalRate	0.20	0.16	0.05	0.29	1.00	-0.26	0.13	-0.53	-0.12	-0.07	0.13
recRate	0.38	0.41	0.46	-0.21	-0.26	1.00	0.49	0.66	-0.05	0.24	-0.37
remPC	0.94	0.59	0.56	-0.35	0.13	0.49	1.00	0.42	0.10	0.35	-0.40
remainRatio	0.29	0.20	0.20	-0.46	-0.53	0.66	0.42	1.00	0.20	0.17	-0.41
newConfDD	0.11	0.11	-0.04	-0.20	-0.12	-0.05	0.10	0.20	1.00	-0.03	-0.31
newDeathDD	0.35	0.55	0.66	-0.37	-0.07	0.24	0.35	0.17	-0.03	1.00	-0.36
newRecDD	-0.29	-0.45	-0.34	0.39	0.13	-0.37	-0.40	-0.41	-0.31	-0.36	1.00

After processing by the Sequential Clustering Dimensionality Reduction Algorithm, the original 11 indices were clustered, dimensionally reduced, and then recombined into the new 7 indices. After that, the correlation of the processed indices was significantly reduced, as shown in Table (3). Multicollinearity between time-series indices was basically excluded, which was more conducive to the analysis and prediction of the following models.

Table 3. Correlation matrix of new data after implementing sequential clustering and dimension reduction algorithm

New index	Index 1	Index 2	Index 3	Index 4	Index 5	Index 6	Index 7
Index 1	1.00	-0.05	-0.10	-0.20	0.06	-0.21	0.10
Index 2	-0.05	1.00	-0.04	-0.05	-0.16	0.36	0.26
Index 3	-0.10	-0.04	1.00	0.22	0.11	-0.14	-0.08
Index 4	-0.20	-0.05	0.22	1.00	0.16	0.09	-0.24
Index 5	0.06	-0.16	0.11	0.16	1.00	-0.11	-0.10
Index 6	-0.21	0.36	-0.14	0.09	-0.11	1.00	-0.08
Index 7	0.10	0.26	-0.08	-0.24	-0.10	-0.08	1.00

2.4.2 Calculation of the relative severity

After dimension reduction, new index $f_{i,j}$ for each country/region was summed by corresponding weights; then Z - Score was selected on cross-sectional statistics to get the relative score of epidemic situation in country *i* at time *t* on each time section. By the Cumulative Sum based on the score on the time series at time section [8], the relative severity $score_{i,t}$ of countries *i* on time-series could be obtained. The formula was as follows:

$$score_{i,t} = \sum_{j=1}^{t} (\beta_1 f_{i,j}^1 + \beta_2 f_{i,j}^2 + \cdots + \beta_n f_{i,j}^n + \varepsilon), t = \{1,2, \dots, T\} \quad (1)$$

Among them, *T* is the time since the outbreak of the epidemic. *n*=7, i.e. 7 new indices.

2.4.3 Regression analysis

Some studies have pointed out that the severity of the epidemic also depends on objective factors in different regions [2]. Therefore, after obtaining the Relative Severity Score of the epidemic, this paper further analyzed the influence of objective conditions on the severity of the outbreak. In this paper, countries with more than 100 confirmed cases were taken as samples, and the following indices were respectively selected for regression analysis on the cumulative confirmed cases, mortality rate and relative severity indices:

Table 4. Factors of regression analysis

Population	Density	Age	Urban Percentage	ICU per 1000	Humidity	Sun Hour	Temp	Wind Speed

Through regression analysis of these geographical, human and social factors, it could be judged whether these factors have significant statistical significance on the severity data of the epidemic. If there is statistical significance, are these factors positively or negatively correlated with the severity of the epidemic?

2.4.4 Future Forecast

Hubei Province, China, South Korea and Japan in the sample are marked as "areas where the epidemic situation has been relatively stable", other countries are marked as "areas where the epidemic is breaking out". In this paper, the Relative Severity Score of the epidemic was input into the supervised machine learning model for regression fitting [9] [10], and the subsequent development of areas where the epidemic is breaking out was predicted by comparing the time-series of areas where the epidemic is relatively stable. The formula was as follows:

$$\widehat{score}_i = f(score_j, \beta) + \varepsilon \quad (2)$$

Where *i* represents the area where the epidemic situation is breaking out, *j* represents the area where the epidemic situation is relatively stable, and f represents the fitting function of the supervised machine learning model.

3 Results

3.1 Relative severity by outbreak countries

Through formula (1), this paper calculated the Relative Severity Score of the epidemic of national/regional samples, as shown in Figure (1).

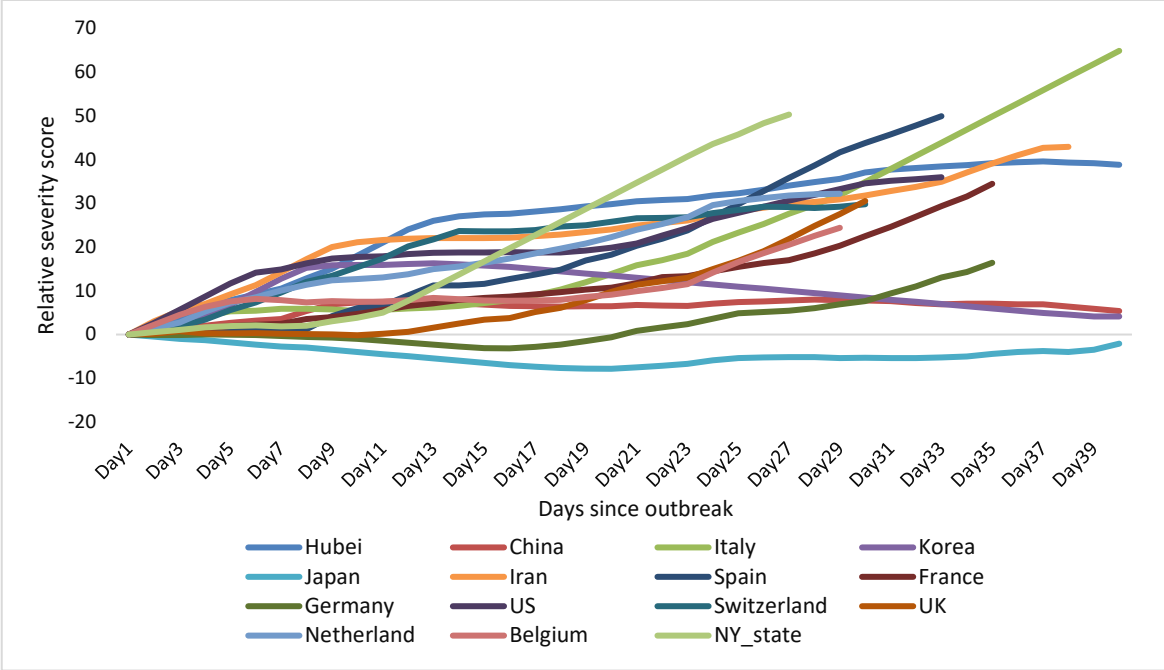

Figure 1. Relative severity score by country/region (part of the samples)

As can be seen from Figure (1), the Relative Severity Score accurately quantified the relative severity of the epidemic in different countries and regions: New York State and Spain became the regions with the fastest outbreak due to the rapid growth and high number of cumulative confirmed cases after the outbreak, surpassing Hubei Province, which was also the hardest hit area of the epidemic, 20 days after the outbreak; Italy has also become the most seriously affected area due to its high mortality rate and rapid growth and large number of confirmed cases. Japan, Germany, South Korea and China have relatively low epidemic severity due to low fatality rates and the rapid control of the number of confirmed cases. However, the recent epidemic situation in Japan has a reversal trend and requires additional vigilance. From the perspective of backtesting, the relative severity of the epidemic measured by the index algorithm was in line with the actual situation, and could compare the epidemic situation in each epidemic area from a deeper level and more dimensions.

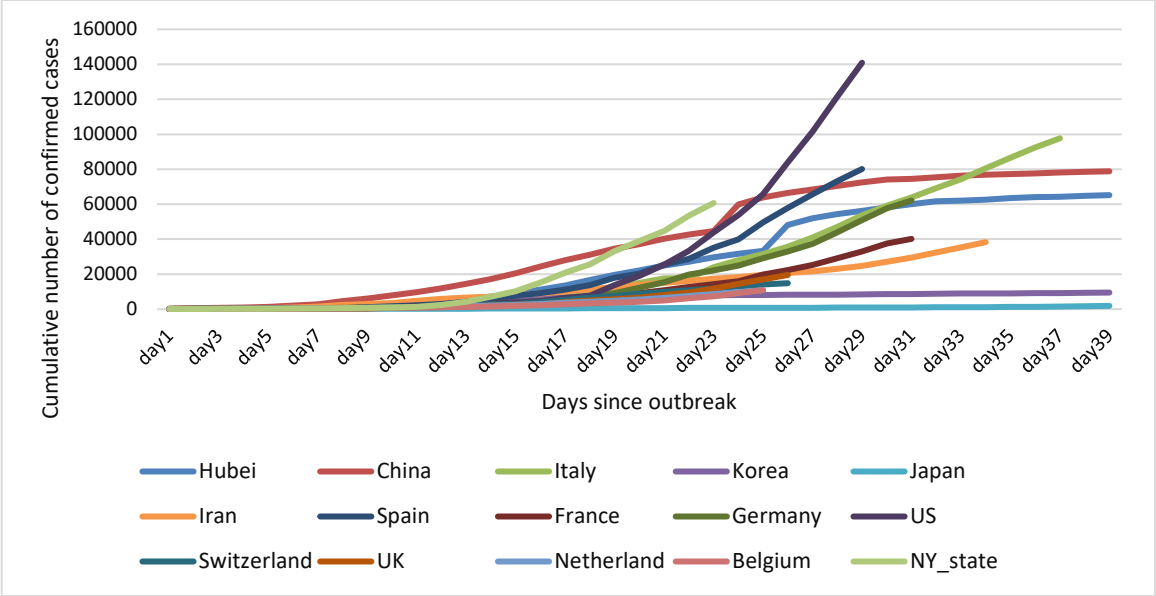

Figure 2. For comparison: Cumulative number of confirmed cases (after reaching 100 confirmed cases)

Figure (2) is the cumulative number of confirmed cases commonly used by the public media and the government when disclosing data[11]. Through the Relative Severity Score map constructed in this paper, it can be found that the cumulative number of confirmed cases map could not well reflect the severity of the epidemic. First of all, the cumulative number of confirmed cases was used in the data, which would face problems such as different population bases and lack of comparability in horizontal comparison. Secondly, as the number of confirmed cases tended to rise exponentially in the outbreak areas[12] [13], the areas with higher number of confirmed cases would completely cover other areas visually, and the image curve was nonlinear and relatively poor in readability, making it difficult to accurately measure the severity of the epidemic. Therefore, the relative Severity score of epidemic situation constructed in this paper can more accurately reflect the trend of epidemic situation in epidemic areas than the traditional cumulative number of confirmed cases.

3.2 Influencing Factors on Outbreak Trend of Epidemic

In this paper, the correlation among the data in Table (4) and the cumulative number of confirmed cases, mortality rate and relative severity score and its corresponding P-value were obtained by regression analysis.

Table 5. Result of regression analysis

Correlation Coefficient	Population [p-value]	Density [p-value]	Age [p-value]	Urban Percentage [p-value]	ICU per 1000 [p-value]	Humidity [p-value]	Sun Hour [p-value]	Temp [p-value]	Wind Speed [p-value]
Confirmed	0.563 **[0.004%]**	-0.046 [73.022%]	0.192 [14.871%]	-0.071 [59.484%]	0.062 [64.643%]	-0.325 **[1.277%]**	0.083 [53.415%]	0.019 [88.661%]	-0.096 [47.160%]
Fatality Rate	0.198 [13.661%]	-0.083 [53.403%]	-0.155 [24.554%]	-0.165 [21.586%]	-0.134 [31.558%]	-0.331 **[1.113%]**	0.183 [16.817%]	0.184 [16.580%]	-0.091 [49.607%]
Relative severity score	-0.147 [27.042%]	-0.033 [80.521%]	0.318 **[1.483%]**	0.231 [8.094%]	0.082 [53.898%]	0.358 **[-0.58%]**	-0.219 [9.937%]	-0.371 **[0.418%]**	0.355 **[0.628%]**

As can be seen from Table (5), the average age, humidity, average temperature and wind speed of the population have significant statistical significance with relatively serious indices. The higher the national average age, average humidity and average wind speed, the more serious the epidemic situation was. The higher the average temperature, the lower the severity of the epidemic.

3.3 Epidemic Forecast

In this paper, through formula (2), the Relative Severity Score of the epidemic situation was taken as input, the data were brought into the machine learning model, and regression analysis was used to fit the data, and the epidemic situation trend in the next month (until May 1, 2020) was predicted.

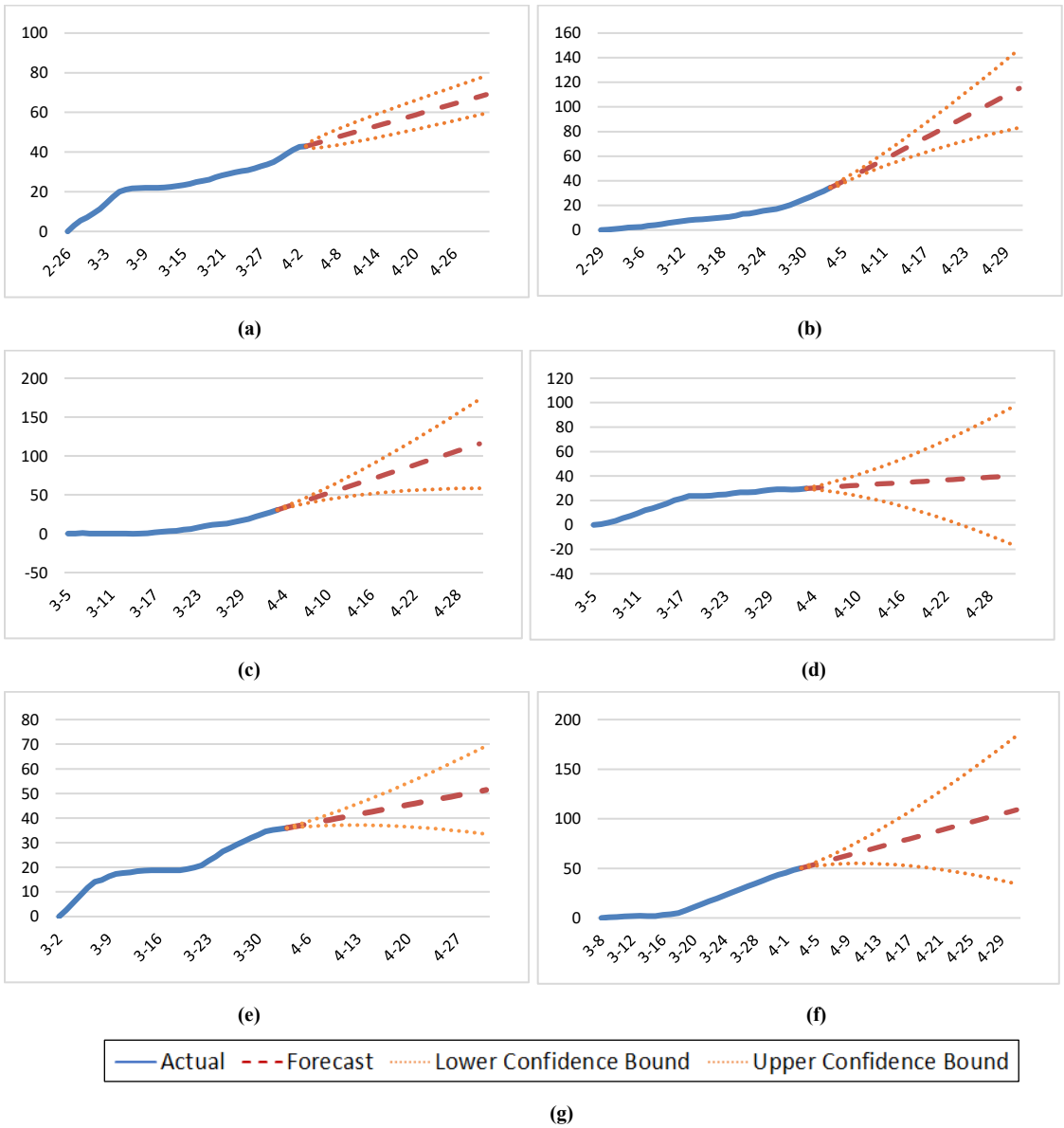

Figure 3. Forecast of Relative Severity Score by country/region (part of the sample)
(a) Iran; (b) France; (c) UK; (d) Switzerland; (e) US; (f) New York state; (g) legends

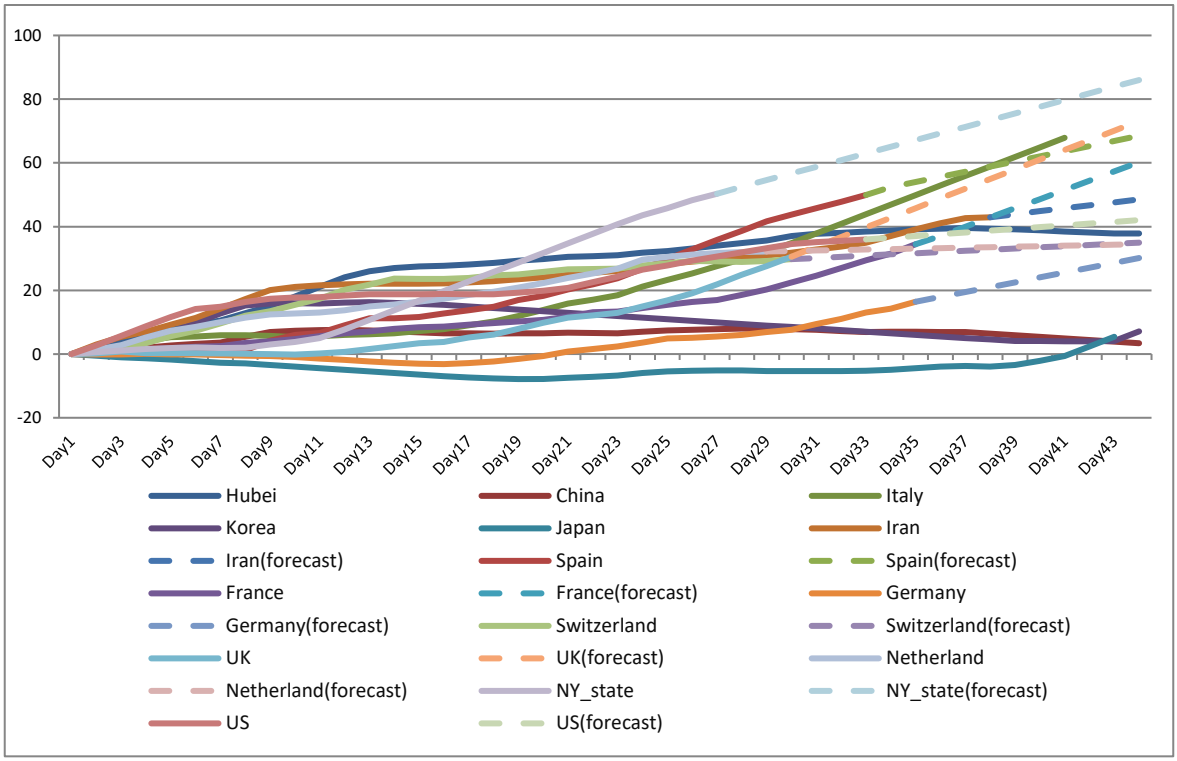

Figure 4. Relative Severity Score and forecast score by country/region (part of the samples)

As can be seen from Figures (3) and (4), after the Relative Severity Score of the epidemic situation was brought into the supervised machine learning model, the model could effectively use the information in the index data to carry out regression analysis and fitting on the future epidemic situation trend. According to the prediction by the model, the epidemic severity in more areas will exceed that in Hubei Province of China in the next period of time. Among them, New York State in the United States will become the relatively serious region, but the trend is also gradually slowing down. Due to the high mortality rate in Italy, there are still a large number of patients who have not been tested, and there is a risk that the medical system will be overwhelmed. The momentum of high growth will continue. Without more measures to intervene, Italy may surpass New York State. The situation in Britain is also worrying. The growth rate of confirmed cases is relatively fast and there is a risk of surpassing Italy. The epidemic situation in Switzerland and Germany has been relatively controlled. According to the forecast, it will usher in an inflection point in the next 2 weeks, with a high degree of similarity with that in Hubei Province.

By analyzing the size of the 95% confidence interval of the prediction results, we could also judge that the the epidemic is developing. For example, Iran and France, as countries with earlier outbreaks of the epidemic, have now taken shape relatively (regardless of the severity of the epidemic), and the fluctuation of relatively serious indices of the epidemic was relatively small, so the confidence interval was also narrow[14]; However, the United States, New York State and Switzerland were regions with short outbreak time. The epidemic situation was still in rapid fermentation, and the relatively serious indices of the epidemic situation fluctuated greatly. Therefore, the confidence interval was wide, and more judgment is needed on the future trend[14].

4 Conclusion

In this paper, a variety of time-series processing methods were used to study the epidemic data of novel coronavirus pneumonia in countries and regions with serious global outbreaks.

Firstly, this paper constructed the Relative Severity Score of the epidemic, The index was constructed by gradually clustering, dimensionality reduction, orthogonalization, regression analysis and other calculations on 11 epidemic data indices such as the cumulative number of confirmed cases per million people in various regions and the growth rate of confirmed cases on that day, excluding the multiple collinear relationships among the data indices, and finally using a linear weighting method. Compared with measuring the severity of the epidemic with a single index such as the cumulative number of confirmed cases and the number of newly confirmed cases, the index added multi-faceted measurement dimensions such as the outbreak speed, the lethality degree of the epidemic, and the recovery speed of the epidemic to more accurately quantify the relative severity of each epidemic area. At the same time, because the index was obtained by linear algorithm, compared with the traditional single index measurement standard in the form of index, the index made the epidemic situation in different regions more intuitive, in line with the actual situation and more readable.

In order to prove the practical application value of this index, this paper also used this index to make regression analysis on some geographical, human and social factors, and found that the average age, humidity, average temperature and wind speed of the population have significant statistical significance with the relatively serious indices. Finally, the index was input into the supervised machine learning model, regression analysis and fitting were carried out on the index, and the follow-up development trend of the epidemic situation in various regions was predicted. Although the prediction results need time to verify, it can be judged that the index data processed by algorithms such as step-by-step clustering and dimension reduction have low signal-to-noise ratio, and the subsequent models using the index as input data all show good stability.

In the future, we will devote ourselves to further improving the relative severity score of the epidemic, adding more dimensional data to the index algorithm, such as the proportion of critically ill patients, the number of cases detected daily (i.e. the daily detection capability of each region), the age of the number of confirmed cases, etc., so that the index can more comprehensively reflect the severity of the epidemic. At the same time, we also hope to add more supervised machine learning algorithms to judge various indices more accurately through neural networks, deep learning and other algorithms, improve the machine learning algorithm, and make the model judge the epidemic trend more accurately.

References

[1] Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). *ArcGIS*. Johns Hopkins CSSE. Retrieved 30 March 2020.

[2] Hannah, Ritchie; Max, Roser (2020). What do we know about the risk of dying from COVID-19?. *Our World in Data*. Archived from the original on 28 March 2020.

[3] Goldberger, Arthur S. (1964). Classical Linear Regression. *Econometric Theory*. New York: John Wiley & Sons. pp. 158.

[4] Ruppert, David. (2010). *Statistics and Data Analysis for Financial Engineering*, Springer Science & Business Media.

[5] Brian, Everitt; Torsten, J Hothorn. (2011). *An Introduction to Applied Multivariate Analysis with R*. Springer, ISBN 978-1441996497.

[6] Zheng, Tanran. (2020). Sequential Clustering and Dimension Reduction Algorithm of Time Series Data. *CEO & CIO in Information Times, ISSN 1007-9440, 23(1): 2-8.*

[7] S. Chatterjee; A.S. Hadi; B. Price. (2000) Regression Analysis by Example (3rd Edition). John Wiley and Sons. ISBN 978-0-471-31946-7.

[8] E. S, Page. (1954). Continuous Inspection Scheme. *Biometrika*. 41 (1/2): 100–115.

[9] Freedman, David. (2009). *Statistical Models: Theory and Practice*. Cambridge University Press. ISBN 978-1-139-47731-4.

[10] Williams, M. N; Grajales, C. A. G; Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research & Evaluation*. 18 (11).

[11] Total confirmed cases of COVID-19 since 100th case. Our World in Data. Retrieved 30 March 2020.

[12] Novel Coronavirus 2019—Situation Updates. WHO. Retrieved 30 March 2020.

[13] Adam, David. (2020). Modelers Struggle to Predict the Future of the COVID-19 Pandemic. *The Scientist Magazine*.

[14] F.M, Dekking. (2005). *A modern introduction to probability and statistics : understanding why and how*. Springer. ISBN 1-85233-896-2.