新冠肺炎疫情时间序列相对严重程度指标及疫情走势预测

郑坦然
（暨南大学产业经济研究院，广东 广州 邮编：510632）

摘　要　由严重急性呼吸道综合征冠状病毒 2 型（SARS-CoV-2）所引起的 2019 冠状病毒（CoVid-19）疫情正在席卷全球。为了使人们更好地了解、评估和控制此次疫情，本文通过构造了一系列衡量疫情严重程度的时间序列数据，并利用逐步聚类及降维算法及最小二乘法回归分析等方法，对疫情已爆发的国家在疫情爆发期内的数据进行了时间序列和横截面的处理，构造了能够准确量化这些国家在疫情爆发后的疫情走势的疫情相对严重程度指标（Relative Severity Score）。在此基础上，通过对该指标和地理、人文和社会因素做回归分析，本文发现人口平均年龄、湿度、平均气温和风速对疫情爆发的相对严重程度具有显著的统计学意义。最后，通过将该指标带入到机器学习模型中使用回归分析对数据进行拟合，本文对疫情正在爆发的国家的未来疫情发展趋势进行了预测。

关键词　新型冠状病毒肺炎；逐步聚类降维算法；时间序列分析；回归分析；机器学习；

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 引言

冠状病毒肺炎（CoVid-19）疫情在世界多个国家爆发。2020 年 2 月 29 日，该疫情被世界卫生组织定义为全球大流行传染病（pandemic）。截止至 3 月 30 日，Covid-19 已造成全球超过 75 万人被感染，35388 人死亡，201 个国家和地区有确诊病例[1]。自疫情爆发以来，已经有文献和研究对疫情的传播和发展进行建模和预测，但由于大部分传染病模型都涉及到对模型参数进行估计，预测准确率有限。因此本文希望研究一个基于疫情真实传播情况对疫情走势的严重性进行评估，并对未来走势进行预测。

经过初步分析发现，虽然疫情在各个国家和地区的爆发有时间差，但疫情走势具有高度相似性，因此本文通过将各个国家和地区在疫情爆发后的疫情数据放在同一时间轴上进行横截面分析，可以量化和比较不同国家和地区在爆发后的疫情严重程度，即疫情相对严重指标（Relative Severity Score）。

疫情相对严重指标可以对比和判断不同国家和地区的疫情发展情况。有研究指出，疫情的严重程度也取决于不同地区的客观因素[2]。因此得到疫情相对严重程度指标后，本文进一步分析地理、人文和社会因素（如平均气温、人口密度、卫生及医疗资源等）对疫情爆发严重性的影响,通过最小二乘法（Ordinary Least Squares, OLS）[3]回归分析，寻找和判断在统计学意义上对疫情传播或严重程度有显著影响的因素。

最后，根据疫情相对严重指标，本文利用机器学习模型中的回归分析（Regression Analysis）对疫情的未来走势进行预测。

2 方法

2.1 样本选择

在样本选择方面，本文为了确保样本间有可比性，选择了 COVID-19 已经开始有传播或输入病例的国家作为样本。根据约翰霍普金斯大学截止至 2020 年 3 月 31 日的确诊人数数据[1]，本文选取了累计确诊人数超过 100 人的国家作为数据样本。在计算疫情相对严重指标时，由于样本中的中国的湖北省和美国的纽约州疫情相对于各自的国家非常严重，是疫情爆发的重点地区，其累计确诊人数一直占其所属国家的确诊人数 40%以上 [1]；并且中美两国国土面积较大，疫区相对分散，因此本文将这两个地区的数据单独列出进行分析。

2.2 数据分析时间段的选取

由于每个国家疫情集中爆发时间不同，为了使疫情爆发情况具有横向比较性，本文统一将确诊人数达到 100 人作为疫情集中爆发的标志，以此日期开始，分析之后 40 天的疫情走势。例如，中国湖北省的确诊人数于 1 月 19 日首次破百，达到 121 人，因此对于湖北省而言，1 月 19 日为 day-1；意大利的确诊人数于 2 月 24 日首次破百，达 132 人，因此 2 月 24 日为意大利的 day-1。通过该方法可以将不同时期爆发的疫情的时间序列数据对齐，在同一横截面上对疫情严重程度进行对比分析。

2.3 构造量化疫情严重程度的指标

[1] 截止至 2020 年 03 月 30 日，数据来源约翰霍普金斯大学

为了对每个国家的疫情情况进行横向对比，需要选取指标对疫情的严重程度进行量化。确诊人数、死亡人数、治愈人数和新增确诊人数是全球披露疫情情况使用的主要数据，然而这三个数据的横截面可比性不高，因为每个国家和地区的人口基数、确诊人数基数都有很大差别，也并没有直接涵盖疫情的爆发速度的信息。因此本文选择根据这三个数据构建了 11 个具有横向可比性的衍生指标，利用这 11 个衍生指标对疫区情况进行横向对比。

表 1. 疫情严重程度衍生指标 [2]
Table 1. Factors of severity of the COVID-19[3]

指标	简称	说明	计算方法
Confirmed Per Million	confPC	每一百万人的累计确诊人数	累计确诊人数/总人口
Confirmed DoD	confDD	日确诊人数增速	（累计确诊人数/前一日累计确诊人数）-1
Death DoD	deathDD	日死亡人数增速	（累计死亡人数/前一日累计死亡人数）-1
Recovery DoD	recDD	日治愈人数增速	（累计治愈人数/前一日累计治愈人数）-1
Fatality Rate	fatalRate	致死率	累计死亡人数/累计确诊人数
Recovery Rate	recRate	治愈率	累计治愈人数/累计确诊人数
Remain Per Million	remPC	每一百万人现存感染人数	（累计确诊人数-累计死亡人数-累计治愈人数）/总人口
Remain Ratio	remainRatio	现存感染人数占累计确诊人数比	（累计确诊人数-累计死亡人数-累计治愈人数）/累计确诊人数
New Confirmed DoD	newConfDD	当日新增确诊人数增速	（当日新增确诊人数/前一日新增确诊人数）-1
New Death DoD	newDeathDD	当日新增死亡人数增速	（当日新增死亡人数/前一日新增死亡人数）-1
New Recovery DoD	newRecDD	当日新增治愈人数增速	（当日新增治愈人数/前一日新增治愈人数）-1

2.4 模型方法

2.4.1 逐步聚类及降维算法

首先，为了保证源数据工整，模型对 2.3 节中介绍的所有衍生指标进行数据预处理，即异常值填充、绝对中位差去极值（Median Absolute Deviation, MAD）[4]和 Z-Score 标准化，确保所有指标免除离群值和错误数值的干扰，并将所有指标化为相同的量纲[5]。由于指标个数较多，且这 11 个指标基本上都由确诊人数、死亡人数和治愈人数这三个原始数据衍生而来，因此指标间也存在一定的多重共线性。因此模型在对数据进行预处理后，利用逐步聚类及降维算法（Sequential Clustering and Dimension Reduction Algorithm）[6]对指标进行聚类和降维，排除指标间的多重共线性，使得后续的回归模型和预测模型能更好地提炼指标的有效信息[7]。

[2] 数据来源：约翰霍普金斯大学，世界银行
[3] Source: Johns Hopkins University CSSE, World Bank

逐步聚类及降维算法是一种处理多维度时间序列的算法，该算法主要通过借鉴 K 临近算法和贪心算法的思想，按一定顺序逐步对时间序列数据进行聚类；得到已聚类分组的数据后，在每一组内进行逐步普通最小二乘法回归分析，得到的残差作为新数据；最后对每组内数据进行加权平均，在每个组别内合成为一组新的数据，最终达到降维的效果[6]。

首先对原始指标的相关性进行分析，通过表（2）可以看到，有个别指标的时间序列相关性大于 0.6，存在较高的多重共线性。如：当日确诊人数增速和当日死亡人数增速的相关性高达 0.92，都反映了疫情爆发的速度；每一百万人的累计确诊人数和每一百万人现存感染人数在疫情集中爆发期间相关性也高达 0.94。这些因子间的多重共线性都需要被处理。

表 2. 逐步聚类降维前的原始指标相关系数矩阵

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

原始指标	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

通过逐步聚类降维算法处理后，原始的 11 个指标被聚类、降维，然后重新组合成了新的 7 个指标。处理后指标的相关性显著降低，见表（3），时间序列指标间的多重共线性基本排除，更有利于后面模型的分析及预测。

表 3. 逐步聚类降维后的新指标相关系数矩阵

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

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

2.4.2 计算相对严重程度

对数据进行降维后，对每个国家/地区的新指标$f_{i,j}$加权求和，之后在横截面统计取 Z-Score，得到每个时间截面上国家 i 在时间点 t 的疫情相对得分。通过将这些时间截面上的得分在时间序列上累计求和（Cumulative Sum）[8]，可以得到时间序列上国家 i 的相对严重程度$score_{i,t}$。公式如下：

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

其中 T 为疫情爆发至今的时间;n=7，即 7 个新指标。

2.4.3 回归分析

有研究指出，疫情的严重程度也取决于不同地区的客观因素[2]。因此得到疫情相对严重程度指标后，本文进一步分析客观条件对疫情爆发严重性的影响。本文以确诊人数超过 100 人的国家作为样本，分别选取了以下指标对累计确诊人数、致死率和相对严重程度指标进行回归分析：

表 4. 回归分析指标
Table 4. Factors of regression analysis

Population	Density	Age	Urban Percentage	ICU per 1000	Humidity	Sun Hour	Temp	Wind Speed
人口总数	人口密度	人口平均年龄	城市化占比	每千人 ICU 病床数	平均湿度	平均日照时长	平均气温	平均风速

通过对这些地理、人文和社会因素指标进行回归分析，可以判断这些因素是否对疫情的严重程度数据有显著统计学意义；若有统计学意义，这些因素与疫情的严重程度是正相关还是负相关。

2.4.4 未来预测

通过将样本中的湖北省、中国、韩国和日本标记为“疫情已相对稳定地区”，其他国家标记为“疫情正在爆发地区”，本文通过将疫情相对严重程度指标输入到机器学习模型中进行回归拟合[9][10]，对比疫情已相对稳定地区的时间序列，对疫情正在爆发地区的后续发展进行预测。公式如下：

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

其中 i 代表疫情正在爆发地区，j 代表疫情已相对稳定地区，f代表机器学习模型的拟合函数。

3 结果

3.1 各疫情爆发国家的相对严重程度

通过公式（1），本文计算出了国家/地区样本的疫情相对严重指标，详见图（1）。

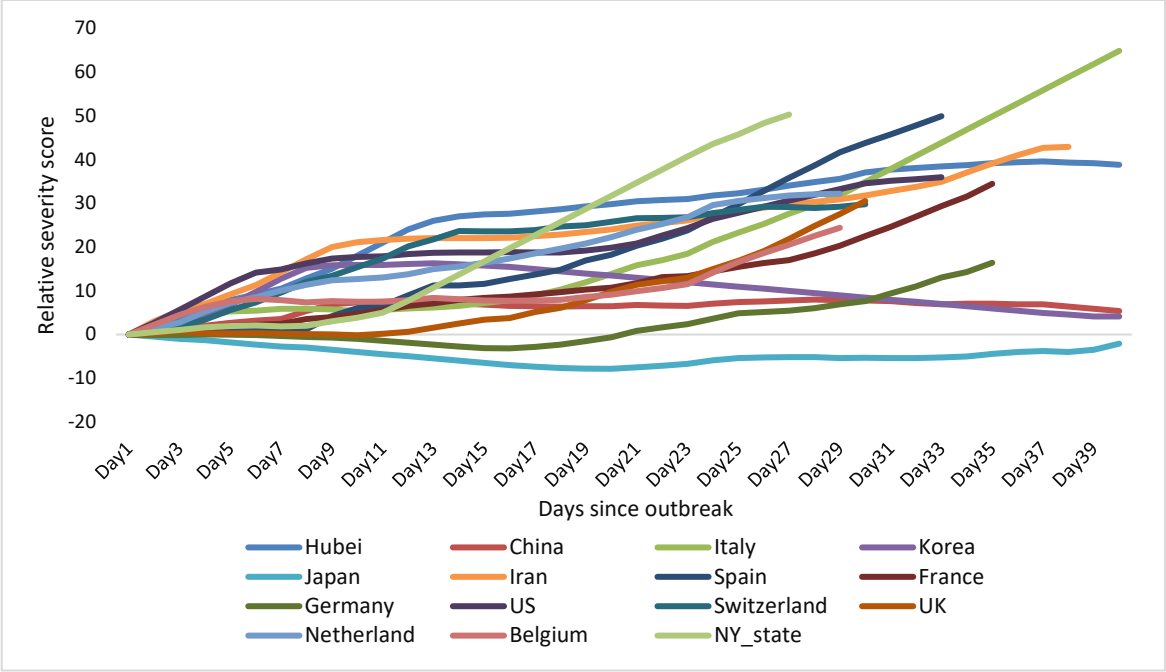

图 1. 各国家/地区相对严重程度指标 （部分样本）

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

从图（1）可以看到，相对严重指标准确地量化了不同国家和地区的疫情相对严重程度：纽约州和西班牙在爆发后由于累计确诊人数快速增长且数量较高，成为疫情爆发最快的地区，在爆发后 20 天就超过了同样是疫情重灾区的湖北省；意大利也因为其致死率高和确诊人数增长较快、数量较多，成为已爆发地区中疫情最严重的地区；日本、德国、韩国和中国因为致死率较低，且确诊人数很快得到控制，疫情严重度相对较低，但其中日本近日的疫情又有反转的趋势，需要额外警惕。从回测的角度看，该指标算法衡量的疫情相对严重程度符合实际情况，能从更深层次、更多的维度对各个疫区的疫情进行对比。

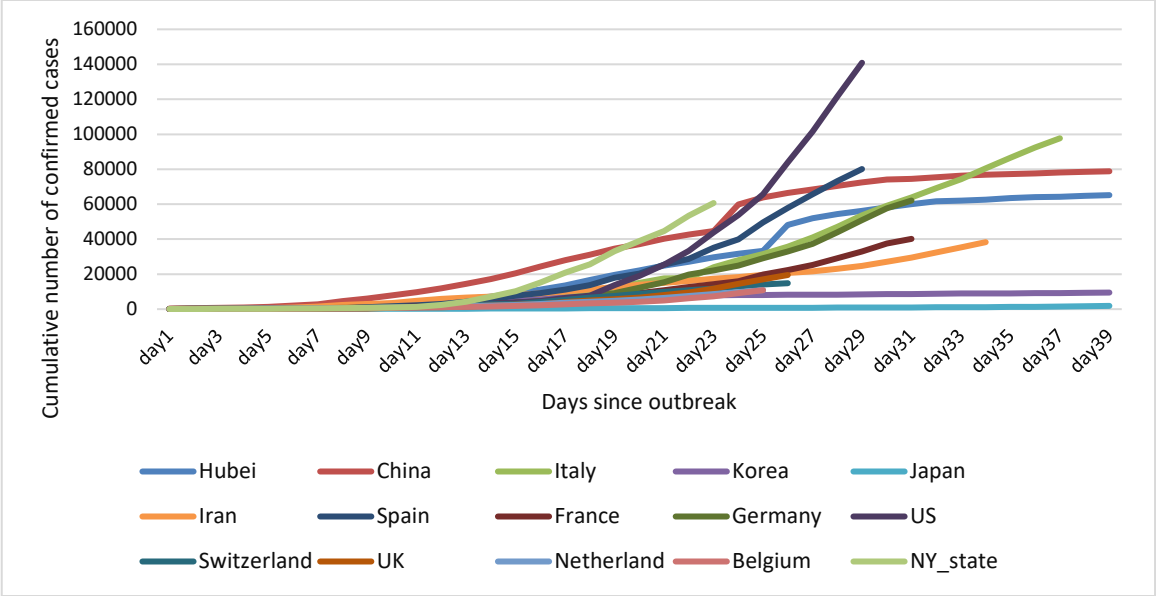

图 2. 对比图：部分国家/地区累计确诊人数（累计确诊人数达 100 人后）[11]

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

图（2）是现在公众媒体和政府公开数据时常用的累计确诊人数图[11]。通过与本文构造的疫情相对严重程度指标图可以发现，累计确诊人数图并不能很好地体现疫情的严重程度。首先数据上采用了累计确诊人数，在横向对比时会面临人口基数不同等问题，缺乏可比性；其次，由于确诊人数在疫情爆发地区往往以指数形式上涨[12][13]，确诊人数较高的地区会在视觉上完全盖过其它地区，图像曲线呈非线性，可读性相对较差、难以准确衡量疫情严重程度。因此本文所构造的疫情相对性严重指标相较于传统的累计确诊人数，更能准确地反应疫区的疫情走势。

3.2 对疫情爆发走势的影响因素

本文通过将表（4）中的数据与累计确诊人数、致死率和相对严重程度指标进行回归分析，得到以下表（5）中的数据间的相关性及其对应的 p-value。

表 5. 回归分析结果

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 **[-0.584%]**	-0.219 [9.937%]	-0.371 **[0.418%]**	0.355 **[0.628%]**

通过表（5）可以看到，人口平均年龄、湿度、平均气温和风速与相对严重指标具有显著的统计学意义。国家的平均年龄、平均湿度和平均风速越大，疫情相对更严重；而平均气温越高，疫情严重性则相对较低。

3.3 疫情预测

本文通过公式（2），将疫情相对严重程度指标作为输入，将数据带入到机器学习模型中使用回归分析对数据进行拟合，对未来一个月（至 2020 年 5 月 1 日）的疫情走势进行了预测。

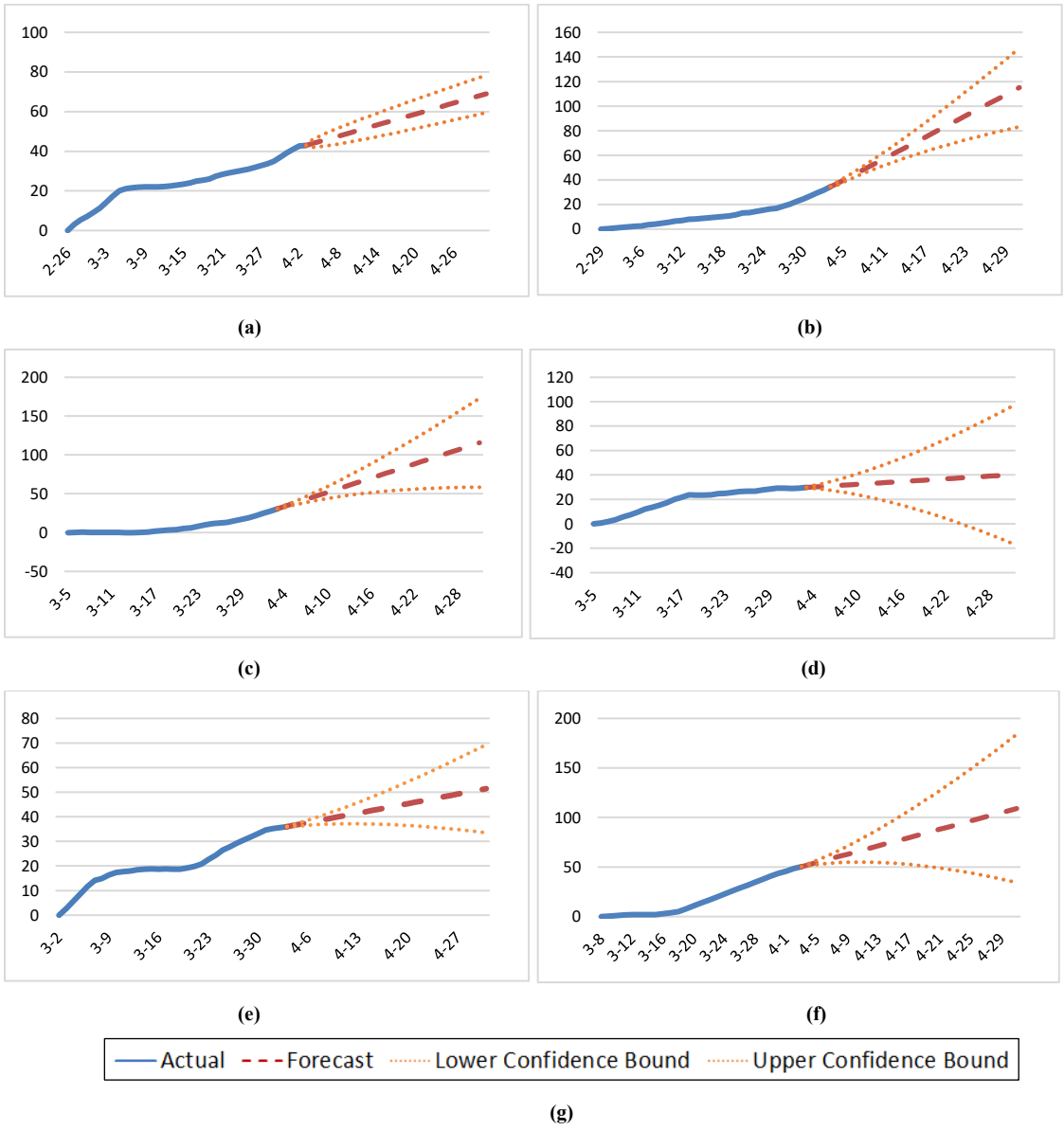

图 3. 后续疫情相对严重指标预测结果（部分样本）

（a）伊朗；（b）法国；（c）英国；（d）瑞士；
（e）美国；（f）美国纽约州；（g）图例

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

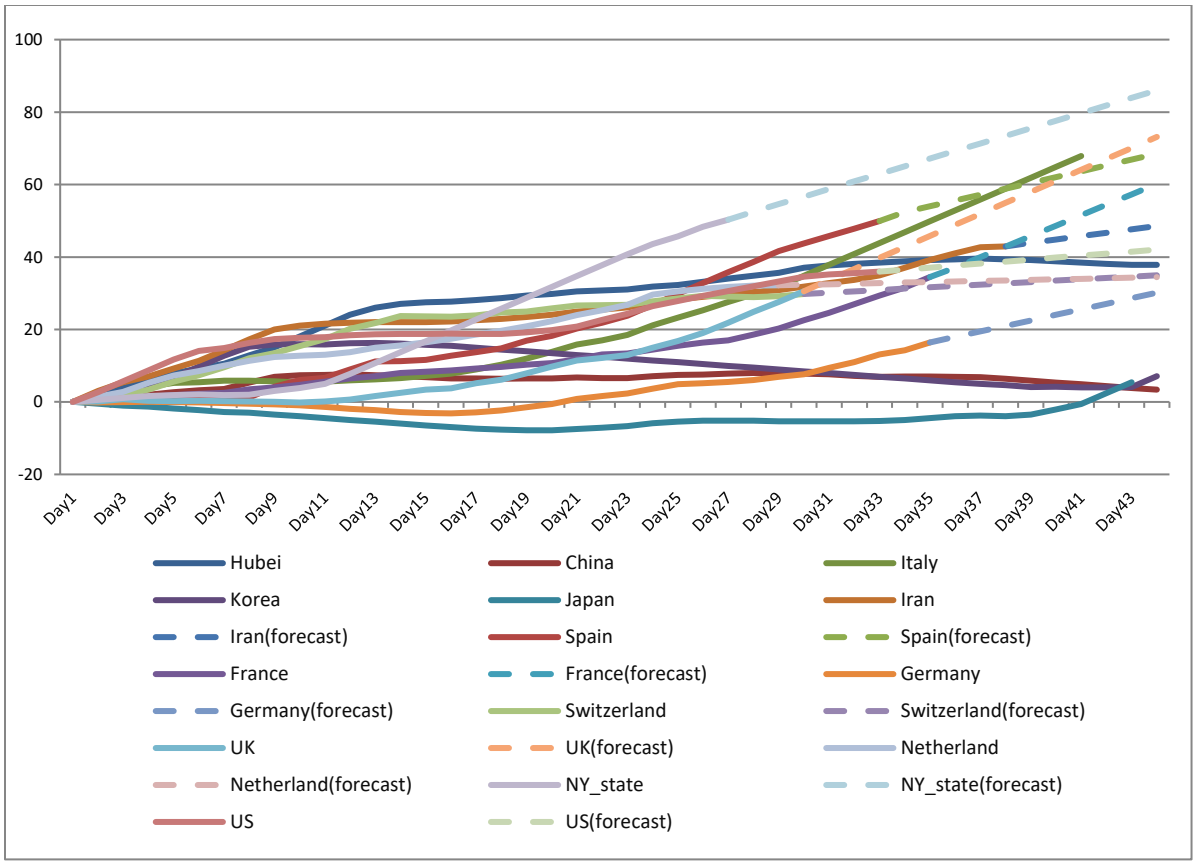

图 4. 各国家/地区相对严重程度指标及其预测值 （部分样本）

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

通过图（3）和图（4）可以看到，将疫情相对严重性指标带入机器学习模型后，模型能够有效的利用指标数据中的信息，对未来疫情走势进行回归分析及拟合。根据模型的预测，接下来一段时间将有更多的地区的疫情严重程度超过中国湖北省；其中美国纽约州将成为相对最严重的地区，但趋势也逐步放缓；意大利由于致死率较高，仍有大量患者未被检测，存在医疗系统不堪重负的风险，高增长的势头或将延续，如没有更多措施介入，有可能超越纽约州；英国的情况也同样堪忧，确诊病例增速较快，有超越意大利的风险；瑞士、德国的疫情已经相对得到控制，根据预测，在未来 2 周内将迎来拐点，走势与湖北省的走势相似度较高。

通过对预测结果的 95%置信区间的大小分析，也可以判断疫情正在发展的情况。例如，伊朗和法国身为疫情较早爆发的国家，现在疫情情况已相对成型（与疫情严重程度无关），疫情相对严重指标的波动较小，因此置信区间也较窄[14]；而美国、美国纽约州和瑞士身为疫情爆发时间较短的地区，疫情仍在快速发酵中，疫情相对严重指标的波动较大，因此置信区间较宽，未来走向需要更多判断[14]。

4 结论

本文利用多种时间序列处理方法，对当下正在爆发的新型冠状病毒肺炎在全球疫情爆发严重的国家和地区的疫情数据进行了研究。

首先本文构造了疫情相对严重程度指标，该指标通过对各地区间每一百万人的累计确诊人数、当日确诊人数增速等 11 个疫情数据指标进行逐步聚类降维、正交化、回归分析等计算，排除各数据指标间的多重共线性关系，最后用线性加权的方式构造而成。相比于用累计确诊人数、新增确诊人数等单一指标衡量疫情严重程度，该指标加入疫情爆发速度、疫情致命程度、疫情恢复速度等多方位的衡量维度，对各个疫区的相对严重程度进行了更准确的量化；同时由于该指标由线性算法得来，因此该指标相对于传统的、呈指数形式的单一指标衡量标准而言，令各地区间的疫情比较更加直观、符合实际情况，可读性更高。

为了证明该指标的实际应用价值，本文还利用该指标对一些地理、人文和社会因素做回归分析，发现人口平均年龄、湿度、平均气温和风速与相对严重指标有显著的统计学意义。最后将该指标输入到机器学习模型中，对该指标进行回归分析及拟合，对各地区的疫情后续发展趋势进行了预测。虽然预测结果需要时间来验证，但可以判断的是，经过逐步聚类降维等算法处理的指标数据，数据信噪比低，利用该指标作为输入数据的后续模型均展现出良好的稳定性。

后续我们将致力于进一步改进疫情相对严重程度指标，在指标的算法中加入更多维度的数据，例如：危重症患者比例、每日病例检测人数（即每个地区的每日检测能力）、确诊人数年龄等，使指标更全面地体现疫情的严重程度。同时也希望加入更多的机器学习算法，通过神经网络、深度学习等算法对各类指标进行更精确地判断，改进机器学习算法，使模型对疫情走势的判断更精确。

参考文献

[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] 郑坦然. 时间序列数据逐步聚类降维法[J]. IT 经理世界, CN11-3928/TN (ISSN1007-9440), 2020, 第 1 期.

[6] Zheng, Tanran. (2020). Sequential Clustering and Dimension Reduction Algorithm of Time Series Data. *CEOCIO China Magazine.*

[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.