

技术方法

DOI:10.13406/j.cnki.cyx.002386

新型冠状病毒肺炎疫情发展区域差异的聚类分析

易大莉¹, 李高明², 冷怀明¹

(陆军军医大学 1. 学报编辑部; 2. 卫生统计学教研室, 重庆 400038)

【摘要】目的:对新型冠状病毒肺炎疫情的地域性差异进行综合评价,对疫情的严重程度进行合理分类。**方法:**采用系统聚类方法对累计确诊病例、累计死亡病例和累计治愈病例的发展趋势进行分类;采用聚类热图进行直观显示和分析。**结果:**第一类:湖北,湖北的确诊人数(56 249 例)远远高于其他地区的总和(12 335 例);第二类:河南,河南确诊人数(1 231 例)处于全国第三位,是湖北以外死亡人数最多(13 例)的地区;第三类:江西、浙江、江苏、上海、四川、安徽、湖南、广东,这些地区确诊人数靠前,病死率较低;第四类:山东、重庆、北京、广西、甘肃、天津、黑龙江、海南、河北,这些地区确诊人数处于居中位置,治愈率和病死率也居中;第五类:山西、云南、福建、陕西、辽宁、贵州、吉林,这些地区确诊人数靠后,死亡人数很少(该类地区合计仅 3 例);第六类:西藏、台湾、青海、宁夏、澳门、内蒙古、香港、新疆,这些地区确诊人数很少(均少于 71 例),死亡人数共 2 例。**结论:**最需要严格控制的地区为湖北和河南特高流行地区,需要密切注意的地区为江西、浙江、江苏、上海、四川、安徽、湖南、广东等高流行地区;西藏、台湾、青海、宁夏、澳门、内蒙古、香港、新疆等地区为相对安全的低流行地区;其他地区属中度流行地区。

【关键词】新型冠状病毒肺炎;疫情;地域性差异;聚类方法**【中图分类号】**R181.8⁺1**【文献标志码】**A**【收稿日期】**2020-02-19

Cluster analysis on epidemic status of COVID-19 among different provinces

Yi Dali¹, Li Gaoming², Leng Huaiming¹

(1. Department of Journal Editorial; 2. Department of Biostatistics, Army Medical University)

【Abstract】Objective: To give the comprehensive evaluation on epidemic status of coronavirus disease 2019 (COVID-19) among different provinces and classify all provinces into reasonable classes in accordance with the severity. **Methods:** Systematic clustering method was used to classify the development trend of cumulative confirmed cases, cumulative death cases and cumulative cured cases. Clustering heat maps were used for visual display and analysis. **Results:** The first class included Hubei. A total of 56 249 confirmed cases in Hubei were much more than the total of 12 335 cases in other regions. The second class included Henan. Confirmed cases of 1 231 in Henan ranked to the third in the country and the death number was the largest (13 cases) except Hubei. The third class included Jiangxi, Zhejiang, Jiangsu, Shanghai, Sichuan, Anhui, Hunan, and Guangdong. These provinces had a higher confirmed cases but lower mortality compared with the following classes. The fourth class included Shandong, Chongqing, Beijing, Guangxi, Gansu, Tianjin, Heilongjiang, Hainan and Hebei. These provinces had a median confirmed patients, cure rate and mortality. The fifth class included Shanxi, Yunnan, Fujian, Shaanxi, Liaoning, Guizhou and Jilin. These provinces had the a lower confirmed patients and very few death cases (only 3 cases in this class). The sixth class included Tibet, Taiwan, Qinghai, Ningxia, Macau, Inner Mongolia, Hong Kong and Xinjiang. These provinces had few confirmed cases (less than 71 cases) and only 2 death cases. **Conclusion:** In summary, areas that need to be closely controlled are Hubei and Henan, where can be regarded as extremely high-endemic areas. Areas that need to be closely monitored are Jiangxi, Zhejiang, Jiangsu, Shanghai, Sichuan, Anhui, Hunan and Guangdong, where can be regarded as high-endemic areas. Tibet, Taiwan, Qinghai, Ningxia, Macau, Inner Mongolia, Hong Kong and Xinjiang are relatively safe and can be regarded as low-endemic areas, while other areas are moderately endemic areas.

【Key words】coronavirus disease 2019; epidemic situation; regional differences; cluster analysis

作者介绍: 易大莉, Email: 413967787@qq.com,

研究方向: 卫生统计学、杂志编辑学。

通信作者: 冷怀明, Email: lenghuaiming@tmmu.edu.cn。

优先出版: <http://kns.cnki.net/kcms/detail/50.1046.R.20200330.1706.008.html>

(2020-03-31)

2019年底,一场突如其来的新型冠状病毒肺炎(coronavirus disease 2019, COVID-19)疫情在中华大地上肆虐,给中国的社会生活、经济生活和精神生活都造成了巨大的冲击^[1]。2020年2月10日,习近平在北京调研指导 COVID-19 疫情防控工作时指出:只有集中力量把重点地区的疫情控制住了,才能从根本上尽快扭转全国疫情蔓延局面^[2]。所谓的“重点地区”是指疫情状况最严重的地区。如何对全国 34 个省市疫情进行科学、合理的分类评价,将“重点地区”筛选出来,这是一个亟待解决的实际问题^[3-4]。本研究将确诊病例、死亡病例、治愈病例作为疫情状况的综合评价指标,采用双向系统聚类热图法^[5-6]对全国 34 个省市的疫情进行分析,评价出高流行、中度流行和低流行区域,从而为国家集中力量控制重点地区的疫情,为防治力量的区域统筹提供科学、合理的理论依据。

本研究从 2 个方向进行聚类:“地区”聚类反映各地区疫情流行程度的差异;“时间点”聚类反映疫情变化的各个阶段,即“上升段”“平稳段”和“下降段”。聚类的主要依据是各地区疫情在所有时间点上“变化趋势”的相似性,而传统方法仅考虑当前时间点的疫情状况(如统计地图法)。本研究将双向系统聚类热图法应用于疫情严重程度的综合评价,以期直观呈现疫情研究程度。

1 数据与方法

1.1 数据来源

本研究数据来源于各省的卫生健康委员会网站^[7-8]。使用 R 语言软件编程,实时抓取网站每日上报的疫情数据,包括累计确诊病例数、累计死亡病例数和累计治愈病例数^[9-10]。

各省疫情数据分布时间始于 2020 年 1 月 11 日,由于 1 月 11 日至 1 月 18 日只有湖北的确诊病例数,且基本维持在 41 例,而其他各省数据始于 2020 年 1 月 19 日。所以,本研究纳入分析数据的时间段为 2020 年 1 月 19 日至 2020 年 2 月 16 日 10:25:00。

1.2 数据整理

本研究纳入分析的指标为累计确诊病例数、累计死亡病例数和累计治愈病例数,数据整理格式见表 1。

其中,每一行数据代表该地区疫情变化趋势(即时间序列)。聚类分析的基本思想是把疫情变化趋势相似的地区聚为一类,从而把全国 34 个省市划分为不同的类别。采用这种

方式的聚类结果能够综合反映疫情的严重程度。

在综合评价分析中,原则上应该采用相对数,即确诊病例、死亡病例和治愈病例数的绝对数除以该地区的总人口数。但是这 3 个指标的绝对数较小(湖北除外),除以全省人口数后,其相对数几乎为 0;且由于该传染病的传染性极强,人们最关注的是疫情的绝对规模。综上所述,本研究在进行聚类分析时采用绝对数。

1.3 方法

1.3.1 聚类分析方法 本研究采用系统聚类法(分层聚类法)。这里,定义聚类的“样品”是指某个疫情指标形成的时间序列(即趋势曲线),代表该疫情指标随时间的发展变化趋势。系统聚类法首先把每个“样品”作为一类,然后把最靠近的样品(相距距离最小,即发展变化趋势最接近)首先聚为不同的小类,再将已聚合的小类按其类间距离进行再合并,这样不断继续下去,最后把所有的子类都聚合到一个大类^[11]。

需要说明的是,由于湖北地区的疫情特别严重,其疫情数据远远高于其他地区,如果和其他地区放在一块儿进行聚类分析,其聚类结果只能分为 2 类。因此,把湖北地区单独作为一类,并在后面的分析结果中标为最严重的“第一类”。

1.3.2 聚类热图与分类数 本研究将系统聚类过程采用聚类热图进行直观表示,并在此基础上确定合理的分类数。聚类热图设计如下:纵轴方向代表不同地区,该地区对应“行”的颜色变化代表疫情变化,不同颜色对应“数据大小”标识在图右侧上方的标尺;横轴方向代表时间点,一个时间点对应“列”的颜色变化代表不同地区疫情的差异。而顶端和左侧的折线簇表示系统聚类的过程^[12-13]。

关于分类数,由于聚类热图的优势是能够直观反映聚类结果,即颜色相近的“行”或“列”应该聚为一类,所以,本研究直接根据颜色变化层次并结合实际判断出合理的分类数。

2 结果

2.1 确诊病例疫情分析

由图 1 所示,确诊病例整体上呈上升趋势,按严重程度可分为 7 类。第一类:湖北;第二类:河南、广东、浙江;第三类:湖南、安徽、江西;第四类:四川、江苏、山东、重庆;第五类:北京、黑龙江、河北、广西、陕西、福建、上海;第六类:海南、云南、甘肃、吉林、辽宁、山西、贵州、天津;第七类:西藏、澳门、青海、台湾、香港、内蒙古、宁夏、新疆。

表 1 各地区病例数时间序列

序号	地区	2020-01-19	2020-01-20	……	2020-02-14	2020-02-15	2020-02-16
01	安徽	xxx	xxx	……	xxx	xxx	xxx
02	澳门	xxx	xxx	……	xxx	xxx	xxx
……				……			
33	浙江	xxx	xxx	……	xxx	xxx	xxx
34	重庆	xxx	xxx	……	xxx	xxx	xxx

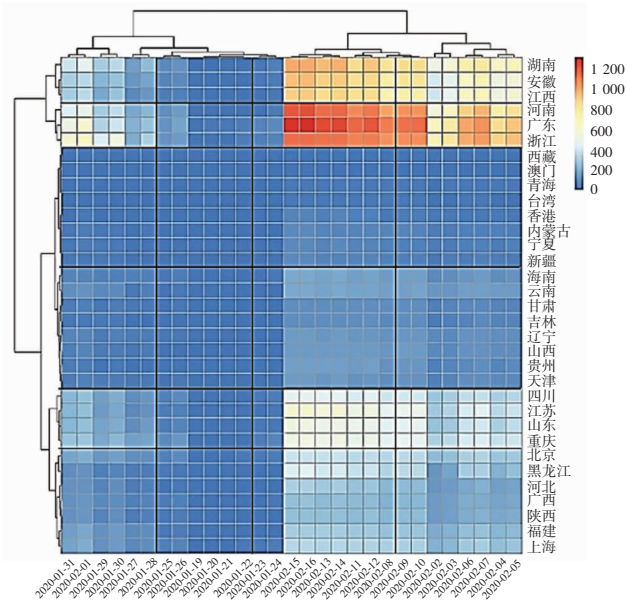


图 1 确诊病例区域聚类结果

时间段的划分主要依据系统聚类折线和颜色变化。小方块颜色相近的划分为一类,对应一个时间段。而要判断这时间段内的变化快慢,则根据小方块颜色色差,色差越大表示变化速度越快,色差越小表示变化速度越慢。根据图 1 分析得出,确诊病例时间变化划分为 4 个阶段:2020-01-19 至 2020-01-24,平稳,出现零星确诊病例;2020-01-25 至 2020-02-01 部分地区病例增加较快;2020-02-02 至 2020-02-07 近一半地区快速上升;2020-02-08 至 2020-02-16 病例数继续上升,但增速减缓。具体见表 2。

2.2 死亡病例疫情分析

死亡病例整体上升,按严重程度可分为 6 类(图 2)。第

一类:湖北;第二类:河南、黑龙江;第三类:安徽、重庆、河北、北京、海南;第四类:天津、甘肃、湖南、广东、广西、山东;第五类:江西、辽宁、新疆、香港、贵州、吉林、上海、四川;第六类:浙江、云南、西藏、台湾、陕西、山西、青海、宁夏、内蒙古、江苏、澳门、福建。

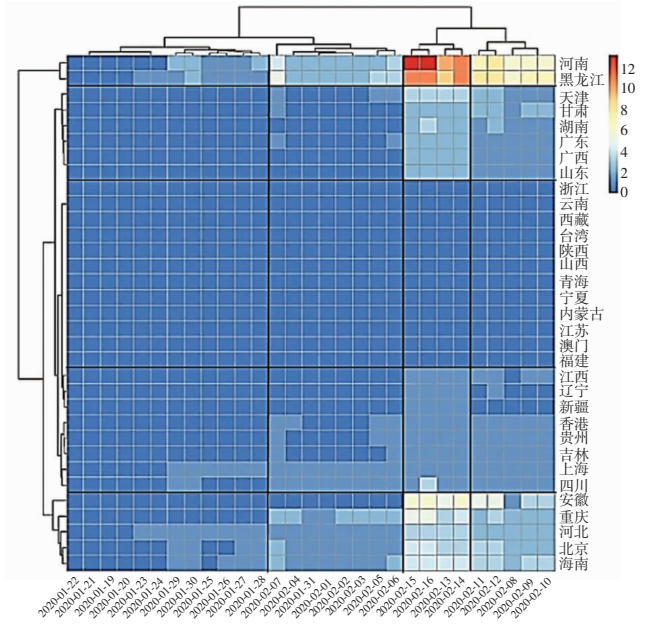


图 2 死亡病例区域聚类结果

由图 2 分析得出,死亡病例时间变化划分为 4 个阶段:2020-01-19 至 2020-01-30,平稳,少数地区出现零星死亡病例;2020-01-31 至 2020-02-07,平稳,部分地区出现个位数的死亡病例;2020-02-08 至 2020-02-12,缓慢上升,近半地区出现死亡病例;2020-02-13 至 2020-02-16,持续增加,有 2 个地区出现 10 例以上死亡病例。具体结果见表 3。

表 2 基于确诊病例在不同时间段的疫情分析

类别	2020-01-19 至 2020-01-24	2020-01-25 至 2020-02-01	2020-02-02 至 2020-02-07	2020-02-08 至 2020-02-16	流行程度
第一类(湖北)	...	...	...	...	特高流行
第二类	病例少平稳	上升快	上升快	缓升增速减	高流行
第三类	病例少平稳	上升较快	上升较快	缓升增速减	高流行
第四类	病例少平稳	缓慢上升	缓慢上升	平稳波动	中度流行
第五类	病例少平稳	缓慢上升	缓慢上升	平稳波动	中度流行
第六类	病例少平稳	病例少平稳	缓慢上升	平稳波动	低流行
第七类	病例少平稳	病例少平稳	缓慢上升	平稳波动	低流行

注:聚类分析中不包括湖北,因此不包括湖北变化趋势。以下同

表 3 基于死亡病例在不同时间段的疫情分析

类别	2020-01-19 至 2020-01-30	2020-01-31 至 2020-02-07	2020-02-08 至 2020-02-12	2020-02-13 至 2020-02-16	流行程度
第一类(湖北)	...	...	...	...	特高流行
第二类	零星死亡	上升快	缓慢上升	上升快	高流行
第三类	零星死亡	缓慢上升	缓慢上升	缓慢上升	高流行
第四类	零星死亡	零星死亡	缓慢上升	平稳波动	中度流行
第五类	死亡无	零星死亡	缓慢上升	平稳波动	低流行
第六类	死亡无	死亡无	死亡无	死亡无	低流行

2.3 治愈病例疫情分析

治愈病例呈现上升趋势,按例数由“多”到“少”可分为 5 类(图 3)。第一类:湖北;第二类:广东、河南、湖南、浙江;第三类:安徽、江西、江苏、山东、重庆、四川、北京、上海;第四类:福建、河北、广西、山西、甘肃、海南、云南、黑龙江、陕西;第五类:内蒙古、青海、新疆、西藏、澳门、台湾、香港、天津、贵州、辽宁、吉林、宁夏。

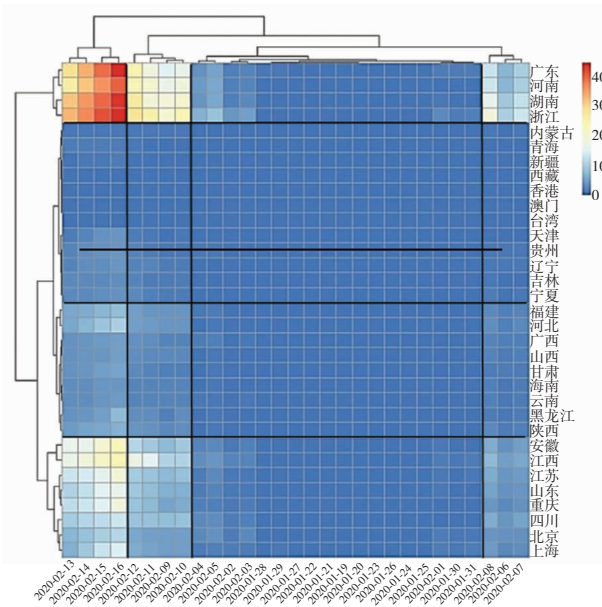


图 3 治愈病例区域聚类结果

根据图 3 分析得出,治愈病例时间变化划分为 4 个阶段:2020-01-19 至 2020-02-05,平稳,出现少量治愈病例;2020-02-06 至 2020-02-08,缓慢上升,4 个地区出现较多治愈病例;2020-02-10 至 2020-02-12,上升,治愈病例上升较快;2020-02-13 至 2020-02-16,上升,治愈病例快速上升。具体结果见表 4。

表 4 基于治愈病例在不同时段的疫情分析

类别	2020-01-19至 2020-02-05	2020-02-06至 2020-02-08	2020-02-10至 2020-02-12	2020-02-13至 2020-02-16
	零星治愈	零星治愈	零星治愈	零星治愈
第二类	零星治愈	上升	上升	上升加速
第三类	零星治愈	缓慢上升	上升	上升
第四类	零星治愈	缓慢上升	缓慢上升	治愈上升
第五类	无治愈	无治愈	零星治愈	治愈上升

2.4 聚类结果综合评价

由图 4 所示,综合评价结果如下。疫情最为严重的特高流行地区是湖北和河南;高流行地区为江西、浙江、江苏、上海、四川、安徽、湖南、广东;中度流行地区为山东、重庆、北京、广西、甘肃、天津、黑龙江、海南、河北、山西、云南、福建、陕西、辽宁、贵州、吉林;低流行地区为西藏、台湾、青海、宁夏、澳门、内蒙古、香港、新疆。

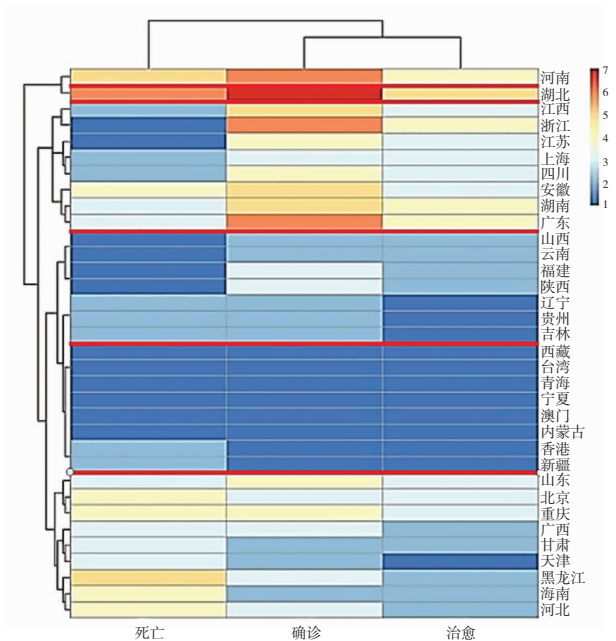


图 4 综合聚类结果热图

3 讨论

3.1 基于确诊病例聚类结果的解读

确诊病例反映疫情流行态势。湖北的确诊人数(56 249 例)远远高于其他地区的总和(12 335 例),“严防输出”是控制整个疫情的重中之重,“内防扩散”也是疫情防控的第一要务。广东、河南、浙江的确诊人数均数为 1 100 例,为高流行区域,应该采取“内防扩散、外防输入、严防输出”的防控策略。湖南、安徽、江西、四川、江苏、山东、重庆、北京、黑龙江、河北、广西、陕西、福建、上海的确诊人数均数为 399 例,为中度流行区域,应该实施“外防输入、内防扩散”的防控策略,对密切接触者进行全面排查和集中隔离医学观察。海南、云南、甘肃、吉林、辽宁、山西、贵州、天津、西藏、澳门、青海、台湾、香港、内蒙古、宁夏、新疆的确诊人数均数为 83 例,为低流行区域,应该执行“外防输入”的防控策略,抓好重点地区流动人员跟踪管理,强化监测发现报告、预检分诊,规范发热门诊等工作,及时开展流行病学调查和密切接触者追踪管理,落实社区防控措施,交通运输、城乡公共交通正常运转。

3.2 基于死亡病例聚类结果的解读

死亡病例从某种程度反映疾病致死的严重程度,因此是医疗力量投入的依据。湖北的死亡人数(1 596 例)远远高于其他地区的总和(70 例),显然是医疗力量投入的重中之重,紧急征调国家救援力量和各省市医疗资源是积极救治和降低病死率的重要举措。河南、黑龙江死亡人数高,均数为 12 例(1.432%),虽然广东和浙江为高流行区域,但死亡病例低。安徽、重庆、河北、北京、海南、天津的平均死亡人数为 4 例(0.078%)。其余地区的死亡病例均数为 0.6 例(0.022%)。

3.3 基于治愈病例聚类结果的解读

治愈病例从某种程度反映该地区的医疗水平,也是地区差异性需要考虑的方面。湖北的治愈人数为 5 623 例(10.00%),但治愈率不高。广东、浙江、湖南虽然确诊人数多,

但治愈率高(37.36%),说明医疗水平高。安徽、江西、江苏、山东、重庆、四川、北京、上海、福建、河北、广西、山西、甘肃、海南、云南、黑龙江、陕西地区的平均治愈人数为 108 例(28.10%),处于中流水平。

3.4 对地区总体聚类结果的解读

在考虑应急分级处理时,不仅要考虑确诊人数,也要考虑当地的医疗力量和水平。总体聚类结果兼顾了这 3 个方面的信息。第一类:湖北,是特高流行地区;第二类:河南,确诊人数(1 231 例)处于全国第三位,但由于较高的病死率(1.06%),被单独聚为一类,也应划分为特高流行地区;第三类:江西、浙江、江苏、上海、四川、安徽、湖南、广东,这些地区确诊人数靠前,但病死率较低,被聚为一类;第四类:山东、重庆、北京、广西、甘肃、天津、黑龙江、海南、河北,这些地区确诊人数处于居中位置,治愈率和病死率也居中,因此被聚为一类;第五类:山西、云南、福建、陕西、辽宁、贵州、吉林,这些地区确诊人数靠后,死亡人数很少(该类地区合计仅 3 例),被聚为一类;第六类:西藏、台湾、青海、宁夏、澳门、内蒙古、香港、新疆,这些地区确诊人数很少(均少于 71 例),该类地区死亡人数共 2 例,因此聚为一类。

综上所述,最需要严格控制的地区为湖北和河南,应该采取严格的防控策略,并通过严格的趋势观察,考虑是否恢复生产生活秩序。需要密切注意的地区为江西、浙江、江苏、上海、四川、安徽、湖南、广东,这些地区可视为高流行区,可以根据疫情态势逐步恢复生产生活秩序;西藏、台湾、青海、宁夏、澳门、内蒙古、香港、新疆可视为低流行区,这些地区相对安全,实行全面恢复正常生产生活秩序;而其他中度流行地区,可以根据分类层次,采取相应防控等级,尽快有序恢复生产生活秩序。

3.5 对防控措施的建议

按照突发公共卫生事件的影响范围、危害程度等,突发公共卫生事件应急响应分为 I 级、II 级、III 级、IV 级 4 个等级。I 级响应是指在发生特别重大突发公共卫生事件时,省级行政区划指挥部根据国务院的决策部署和统一指挥,组织协调本行政区域内应急处置工作。现在全国除港澳台以外,31 个省市都启动了“一级响应”。这对疫情的控制起到了非常积极的作用。

从聚类分析结果来看,不同地区的疫情严重程度差异较大。各地区在“一级响应”的基础上,应适时分析疫情的严重程度与地区差异,科学采取差异化防控措施,以尽量减少疫情对社会经济、人民群众生活带来的负面影响,以达到“疫情严重的地区要集中精力抓好疫情防控工作,其他地区要在做好防控工作的同时统筹抓好改革发展稳定各项工作,特别是要抓好涉及决胜全面建成小康社会、决战脱贫攻坚的重点任务”^[14-15]。

3.6 本研究的优劣势

优势包括:第一,现有的聚类分析和统计地图法往往对一个指标在一个时间点的数据进行聚类,而本文针对时间序列数据聚类,不仅反映绝对值数据大小,同时还包含数据变化趋势的信息;第二,采用双向聚类方法,不仅可以对地区聚类,同时可以反映各指标变化的不同阶段。

劣势包括:第一,本研究聚类结果只反映不同地区的严重程度,从而提出需要重点控制的区域,但是不能给出具体措施;第二,分类数的确定主要根据颜色的大体区别,具有一定的主观性。

参 考 文 献

- [1] Lu R, Zhao X, Li J, et al. Genomic characterization and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding[J]. *Lancet*, 2020, 395(10224): 565–574.
- [2] 习近平. 在中央政治局常委会会议研究应对新型冠状病毒肺炎疫情工作时的讲话[N/OL]. (2020-02-03)[2020-02-19]. http://www.mohrss.gov.cn/SYrlzyhshbzb/dongtaixinwen/SYtupianxinwen/202002/t20200217_359995.html.
- [3] 中国疾病预防控制中心新型冠状病毒肺炎应急响应机制流行病学组. 新型冠状病毒肺炎流行病学特征分析[J]. *中华流行病学杂志*, 2020, 41(2): 145–151.
- [4] Read JM, Bridgen JRE, Cummings DAT, et al. Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions[J]. *medRxiv*, 2020[Epub ahead of print]. DOI: 10.1101/2020.01.23.20018549v2.
- [5] Kim C, Lee H, Devaraj V, et al. Hierarchical cluster analysis of medical chemicals detected by a bacteriophage-based colorimetric sensor array[J]. *Nanomaterials(Basel)*, 2020, 10(1): E121.
- [6] Zolfaghari F, Khosravi H, Shahriyari A, et al. Hierarchical cluster analysis to identify the homogeneous desertification management units[J]. *PLoS One*, 2019, 14(12): e0226355.
- [7] Yu GC. Query information of Wuhan coronavirus infected stats[N/OL]. (2020-02-16)[2020-02-16]. <https://github.com/GuangchuangYu/>.
- [8] 卫生应急办公室. 中华人民共和国国家卫生健康委员会疫情通报[EB/OL]. (2020-02-16)[2020-02-16]. <http://www.nhc.gov.cn/xcs/yqtb/202002/4a1b1ec6c03548099de1c3aa935d04fd.shtml>.
- [9] (美) Dan Toomey 著. 数据科学: R 语言实战[M]. 刘丽君, 李成华, 卢青峰, 译. 北京: 人民邮电出版社, 2017: 83.
- [10] 姜晓东. R 语言绘制中国地图, 并展示流行病学数据[N/OL]. (2018-03-13)[2020-02-16]. <https://cloud.tencent.com/developer/article/1058446>.
- [11] Chong FS, O'Sullivan MG, Kerry JP, et al. Understanding consumer liking of beef using hierarchical cluster analysis and external preference mapping[J]. *J Sci Food Agric*, 2020, 100(1): 245–257.
- [12] You C, Lin QS, Zhou XH. An estimation of the total number of cases of NCIP (2019-nCoV) – Wuhan, Hubei Province, 2019–2020[J]. *China CDC Weekly*, 2020, 6(2): 87–88.
- [13] Jauhainen S, Pohl AJ, Äyrämö S, et al. A hierarchical cluster analysis to determine whether injured runners exhibit similar kinematic gait patterns[J]. *Scand J Med Sci Sports*, 2020[Epub ahead of print]. DOI: 10.1111/SMS.13624.
- [14] 新华社. 习近平主持中央政治局常委会会议研究加强疫情防控工作[N/OL]. (2020-02-04)[2020-02-19]. <http://www.liguoeu.com/start/work/123433.html>.
- [15] 新华网. 《求是》杂志发表习近平总书记重要文章《在中央政治局常委会会议研究应对新型冠状病毒肺炎疫情工作时的讲话》[N/OL]. (2020-02-15)[2020-02-19]. <http://www.nhc.gov.cn/xcs/fkdt/202002/53ae72df52324935b8cfc47e8556bfb1.shtml>.

(责任编辑: 冉明会)