

技术方法

DOI: 10.13406/j.cnki.cyxh.001369

基于时序特征的手足口病疫情模式聚类研究

周 亮, 陈虹汝, 李高明, 张彦琦, 伍亚舟, 易大莉, 易 东, 刘 岭

(陆军军医大学军事预防医学系卫生统计学教研室, 重庆 400038)

【摘要】目的: 根据全国 31 个地区 2010 至 2015 年手足口病发病的时间序列特征, 划分不同的疫情模式, 为该疾病防控的统一规划和实施提供科学依据。**方法:** 提取各地区手足口病发病率的时间序列特征, 包括季节波动、趋势、自相关和混沌等; 通过层次聚类分析, 划分不同的时间序列疫情模式。**结果:** 聚类分析共产生 3 类疫情模式, 分别为非季节模式、单发季节模式和多发季节模式。**结论:** 对非季节模式地区, 需随时加强疾病监控, 进一步分析其相关因素; 对单发季节模式和多发季节模式, 仅需在周期性高发来临时作好相应的预防措施, 从而降低疾病大面积流行的可能性。

【关键词】 手足口病; 时间序列; 时序特征; 聚类分析; 疫情模式

【中图分类号】 R183.1

【文献标志码】 A

【收稿日期】 2017-01-29

Cluster analysis of epidemic characteristics of hand-foot-and-mouth disease based on temporal characteristics

Zhou Liang, Chen Hongru, Li Gaoming, Zhang Yanqi, Wu Yazhou, Yi Dali, Yi Dong, Liu Ling

(Teaching and Research Section of Health Statistics, College of Preventive Medicine,
Army Military Medical University)

【Abstract】Objective: to provide a scientific basis for the prevention and control of the hand-foot-and-mouth disease (HFMD) by dividing different epidemic models according to the time series characteristics in 31 regions from 2010 to 2015. **Methods:** Time sequence features of HFMD incidence in different regions was extracted, including seasonal fluctuations, trends, autocorrelation and chaos, etc. Different time series epidemic patterns were divided by hierarchical clustering analysis. **Results:** Cluster analysis produced three types of epidemic patterns, namely, non-seasonal mode, single-season and multi-season seasonal pattern. **Conclusion:** For non-seasonal patterns, disease surveillance should be strengthened at any time to further analyzing the relevant factors. In the single-season and multi-season model, the appropriate preventive measures should be taken only in the high incidence period to reduce the possibility of a large area of epidemic disease.

【Key words】 hand-foot-and-mouth disease; time series; timing characteristics; cluster analysis; epidemic model

手足口病(hand-foot-and-mouth disease, HFMD)是由于广谱的肠道病毒病原体引起的一种常见儿童疾病^[1]。研究显示, HFMD 的发病趋势呈现周期性, 例如在马来西亚、日本、我国台湾地区等地的 HFMD 就存在两年一次的高发或者隔 1 年或 2 年出现一次大规模暴发^[2]。目前, 由于病毒类型、气象因素以及人口学因素的影响, 不同地区 HFMD 的疫情在时

空范围内呈现出不同的发病趋势^[3]。例如在我国台湾地区、香港地区和新加坡等地, HFMD 的发病高峰有 2 个, 分别为 5 至 7 月和 10 至 11 月^[4]; 而另一项来自日本 2007 年的数据显示, 其发病高峰为 8 月和 9 月^[5]; 越南 2005 年的数据显示, 其发病高峰分别为 3 至 5 月和 10 至 11 月^[6]。

目前, 研究认为有效的疾病监测与预警、准确及时地调查与处理疫情, 是控制疾病大规模流行的有效措施^[7]。在前期的研究中, 研究者大多根据 HFMD 发病因素的不同来划分发病模式^[3]; 但由于研究过程中所纳入的因素有限, 或研究之间纳入的因素存在一定差异, 从而导致研究结果的可靠性有待商榷。

随着大数据的研究理念深入人心, 人们的关注

作者介绍: 周 亮, Email: 13983081938@163.com,

研究方向: 流行病与卫生统计学。

通信作者: 刘 岭, Email: liuling_505@sina.com。

基金项目: 国家自然科学基金资助项目(编号: 81473068); 国家社会科学基金资助项目(编号: 14BTJ019)。

优先出版: <http://kns.cnki.net/kcms/detail/50.1046.R.20170630.1039.028.html>
(2017-06-30)

重心逐渐由事物间的因果关系过渡到数据本身的关联性上^[8]。从数据分析的角度可以将手足口病的发病情况以一条时间序列来呈现。本次研究中,可以通过分析不同地区疾病发生的时间序列特点,探寻其相似性,进而划分发病模式。但由于时间序列的高维性,传统的聚类方法容易受到噪声影响,导致聚类结果意义不明确,特别是在缺失数据或不等长序列存在的情况下,聚类结果的有效性会更差^[9]。鉴于此,一些研究提出了基于时序特征的聚类方法,即通过提取时序数据的特征参数(周期强度、趋势性、季节波动等)构建特征向量,从而达到对数据进行有效降维,以保证聚类结果的有效性^[10]。因此,本次研究目的是基于我国不同地区 HFMD 发病的时序特征进行聚类分析,划分不同的疾病发病模式,为该疾病防控的统一规划和实施提供科学依据。

1 资料与方法

1.1 资料来源

本次研究资料来源于中国公共卫生科学数据中心,包括全国 31 个地区 2010 年 1 月至 2014 年 12 月登记的 HFMD 数据(含发病人数、死亡人数、发病率、死亡率等)。

1.2 方法

1.2.1 时间序列特征向量 由于 HFMD 的发病特点呈现明显的周期性,为了探索不同地区时间序列潜在的相似变化机制和结构,本次研究考虑的主要特征为 HFMD 序列数据的频域特征(周期强度)和时域特征(趋势性、季节波动、Box-Pierce 指数、Hurst 指数和李雅普诺夫指数),所有参数的运算可根据 Wang 编写的 R 程序实现^[10]。

1.2.2 层次聚类分析 将上述提及的参数合并在一起以构建 HFMD 时序数据的特征向量,根据各地区数据中特征向量的差异进行层次聚类分析;该方法可以通过 R 程序中的 Cluster 包和 NbClust 包实现;具体实施步骤如下:①数据标准化:采用 Z-Score 法进行数据标准化;②计算距离:通过 $\text{dist}(x, \text{method})$ 函数计算欧式距离,其中 x 为表示输入数据,method 为距离测定方法,默认为欧几里得距离;③确定类别个数:通过 $\text{NbClust}(x, \text{distance} = , \text{min.nc} = , \text{max.nc} =)$,其中 distance 为数据间距离情况,min.nc 为最小聚类数,max.nc 为最大聚类数;④聚类分析:通过 $\text{hclust}(d, \text{method})$ 函数实现聚类,其中 d 为函数 dist 产生的距离矩阵,method 为聚类实现的具体方法,本次研究用到的是 Ward 聚类法(两类间的距离定义为它们之间所有变量的方差分析的平方和)。

2 结果

2.1 时间序列特征向量构建

通过计算获取了全国 31 个地区的 HFMD 时序数据的特征参数,并将每个地区的参数组合在一起,构成该地区的时

序数据的特征向量,详见表 1。

表 1 时间序列特征向量
Tab.1 The feature vector of time series

地区	周期强度	趋势性	季节波动	Box-Pierce 指数	Hurst 指数	李雅普诺夫指数
安徽	0.12	0.21	0.20	0.43	0.97	0.57
北京	0.11	0.51	0.95	0.63	0.97	0.56
福建	0.11	0.52	0.83	0.34	0.95	0.58
甘肃	0.11	0.51	0.96	0.44	0.95	0.55
广东	0.10	0.25	0.46	0.59	0.97	0.58
广西	0.07	0.49	0.15	0.32	0.96	0.60
贵州	0.07	0.49	0.10	0.36	0.96	0.59
海南	0.07	0.42	0.09	0.24	0.93	0.60
河北	0.11	0.56	0.95	0.59	0.97	0.55
河南	0.09	0.27	0.03	0.45	0.97	0.59
黑龙江	0.11	0.46	0.96	0.56	0.96	0.56
湖北	0.05	0.68	0.80	0.47	0.96	0.64
湖南	0.07	0.44	0.07	0.38	0.96	0.59
吉林	0.09	0.36	0.03	0.53	0.97	0.60
江苏	0.05	0.66	0.68	0.43	0.96	0.63
江西	0.07	0.50	0.11	0.35	0.96	0.61
辽宁	0.11	0.53	0.98	0.52	0.95	0.54
内蒙古	0.11	0.49	0.98	0.52	0.94	0.55
宁夏	0.11	0.45	0.97	0.61	0.96	0.54
青海	0.09	0.32	0.06	0.43	0.98	0.59
山东	0.11	0.71	0.98	0.63	0.98	0.56
山西	0.11	0.29	0.96	0.52	0.96	0.55
陕西	0.11	0.52	0.96	0.49	0.97	0.55
上海	0.13	0.11	0.06	0.42	0.96	0.56
四川	0.05	0.84	0.80	0.46	0.95	0.64
天津	0.11	0.36	0.95	0.50	0.96	0.54
西藏	0.00	0.41	0.00	0.47	0.95	0.94
新疆	0.11	0.51	0.92	0.48	0.94	0.55
云南	0.12	0.38	0.24	0.41	0.97	0.56
浙江	0.13	0.19	0.06	0.39	0.97	0.56
重庆	0.05	0.80	0.88	0.53	0.95	0.65

2.2 标准化与距离计算

采用 Z-Score 法对各地区特征向量进行标准化,并通过向量间的欧式距离建立向量间的距离矩阵,详见表 2。

2.3 选择最佳聚类数

在计算获取初始聚类结果后,通过 R3.3.2 中的 NbClust () 函数,对最佳类别个数进行“投票”,结果认为分为 2 类的“票数”最高(6),其次是 3 类(4)和 4 类(4),如图 1 所示。

以趋势参数为横坐标、季节波动为纵坐标构建散点图,如图 2 所示。认为各地区序列趋势的区分程度不如季节波动,因此划分类别以季节波动情况为主;再结合实际情况分析,认为本次聚类结果划分为 3 类相对合理。

2.4 层次聚类分析

根据前面的分析结果,将 31 个地区的 HFMD 时序数据划分为 3 类发病模式,包括非季节模式,即发病趋势不受季节变化影响或影响较小,非季节模式的地区包括上海、安徽、浙江、吉林、云南、河南、青海、海南、湖南、贵州、广西和江西;季节单发模式,即发病趋势受季节变化影响,且 1 年高发 1 次,单发季节模式的地区包括福建、内蒙古、甘肃、新疆、山

表 2 欧式距离矩阵
Tab.2 The matrix of euclidean distance

地区	安徽	北京	福建	...	新疆	云南	浙江
安徽	0.00	11.82	11.11	...	14.31	1.11	0.43
北京		0.00	16.38	...	11.89	9.76	16.12
福建			0.00	...	3.38	9.00	11.88
...				0.00	...	...	...
新疆				...	0.00	13.25	16.25
云南						0.00	1.84
浙江							0.00

西、天津、辽宁、陕西、黑龙江、宁夏、山东、广东、北京和河北；季节多发模式，即发病趋势受季节变化影响，且 1 年高发多次，季节多发模式的地区包括西藏、四川、重庆、湖北、江苏；如图 3 所示。

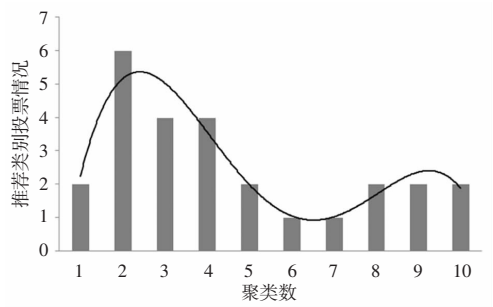


图 1 最佳聚类个数推荐图

Fig.1 The optimal clustering number

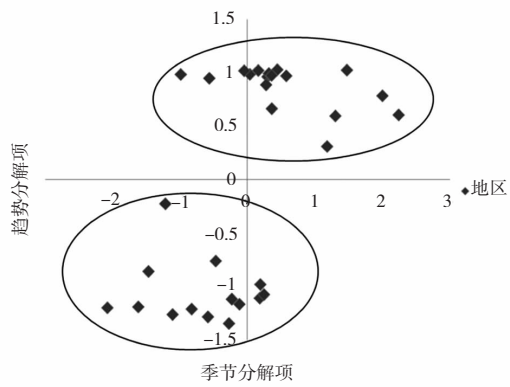


图 2 趋势季节分解项散点图

Fig.2 The scatter diagram of trend and the seasonal decomposition

3 讨论

HFMD 是亚太地区最常见的传染病之一，曾在该地区的多个国家发生大规模流行，引起了世界各国的高度关注^[11-12]。有研究认为 HFMD 一般以 2~3 年为一个周期引起地区性流行，但由于发病因素的差异，不同地区的流行周期存在差异^[9]。本次研究以

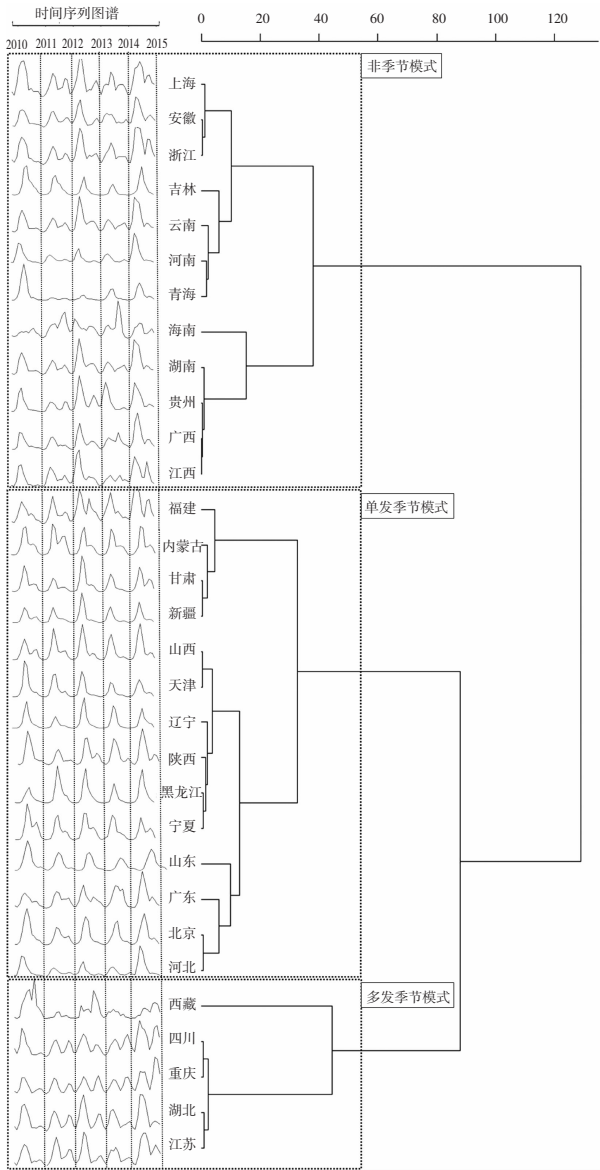


图 3 层次聚类分析结果

Fig.3 The results of hierarchical cluster analysis

时间序列数据的特征向量为基础，将 2010 至 2014 年全国 31 个地区 HFMD 月发病率情况划分出了不同的疫情模式，对统一防控措施制定具有参考价值。

时间序列聚类分析中，由于时序数据存在高维

度、自相关和噪声明显的特点,该类数据聚类时往往不能翻译其真实的数据特征,从而直接导致聚类结果的实际应用价值降低。本次研究采用 Wang 等的方法^[13-14],提取时序数据的特征参数,以各参数构建的特征向量作为分类依据进行分类。结合 HFMD 的发病特点,选择了 5 项参数作为本次聚类的特征向量。其中,周期强度是将时间序列看成多频振荡序列时,反映其频域特征的主要参数^[15];趋势描述时序数据长期变化状态;季节波动反映数据周期内的波动情况;Box-Pierce 指数和 Hurst 指数则是对时序数据自相关性的状态评估;李雅普诺夫指数反映时序数据的混沌性,代表其受初始值的敏感程度^[6]。因此,本次聚类分析兼顾了时序数据的全局性和局部性,从而有效地保证了聚类结果的合理性。

结合软件分析和实际情况,将聚类结果划分为 3 类。其中上海、安徽等地为非季节模式,即这些地区每一年的发病和流行情况季节波动不一致;如上海在 2010 至 2011 年间出现 1 次高峰,而 2011 至 2012 年间则出现 2 次高峰;又如青海在 2010 至 2011 年间出现 1 次高峰,2011 至 2012 年间出现平缓趋势,而 2013 至 2015 年再次出现每年 1 次的小高峰。有研究认为,上海 HFMD 的发病受外来流动人口的影响较大^[16],其可能由于易感人群不断发生变化,从而导致季节波动性不强。与上海相同,青海的主要发病地区主要分布在西宁及海东地区,该地区人口密集、交通经济发达,发病率明显高于其他地区^[17];因此,人口学因素可能是影响非季节模式地区 HFMD 发生的主要因素,这些地区的疾病监控需加强,同时应进一步分析其相关因素,从而提供更为有效的手足口病防控措施。另外,部分地区季节波动性相对明显,因此划分为季节模式。根据发病高峰次数的不同,又将季节模式分为单发模式和多发模式,即每年 1 发和每年多发;如有研究认为广东、福建等地每年的 4 至 7 月会出现比较明显的发病高峰^[18],其主要原因是热带气旋的登陆,对病毒繁殖周期的影响,进而引起疾病在该时间段流行。而四川、重庆等地,发病一般呈现 1 年两发,主高峰在 4 至 7 月,次高峰在 10 至 12 月,其原因可能是由于南方冬季气温高、温度湿度大、适应病毒生长和传播^[19];因此,这些地区需在周期性高发来临时作好相应的预防措施,从而降低疾病大面积流行的可能性。

因此,本次研究基于时序数据特征,对 HFMD 的发病模式进行探索,其结果有实际应用价值,可以为各类别下地区疾病的统一防控规划和对策、资源的优化配置提供合理的参考依据。

参 考 文 献

- [1] Wang Y, Feng Z, Yang Y, et al. Hand, foot and mouth disease in China: patterns of spread and transmissibility[J]. *Epidemiology*, 2011, 22(6): 781-792.
- [2] Podin Y, Gias EL, Ong F, et al. Sentinel surveillance for human enterovirus 71 in Sarawak, Malaysia: lessons from the first 7 years[J]. *BMC Public Health*, 2006, 6: 180.
- [3] 肖革新. 手足口病时空分析及影响因素研究[D]. 北京: 北京师范大学, 2012.
- [4] Chen KT, Chang HL, Wang ST, et al. Epidemiologic features of hand-foot-mouth disease and herpangina caused by enterovirus 71 in Taiwan, 1998-2005[J]. *Pediatrics*, 2007, 120(2): e244-252.
- [5] Miyazawa I, Azegami Y, Kasuo S, et al. Prevalence of enterovirus from patients with herpangina and hand, foot and mouth disease in Nagano Prefecture, Japan, 2007[J]. *Jpn J Infect Dis*, 2008, 61(3): 247-248.
- [6] Tu PV, Thao NT, Perera D, et al. Epidemiologic and virologic investigation of hand, foot, and mouth disease, southern Vietnam, 2005[J]. *Emerg Infect Dis*, 2007, 13(11): 1733-1741.
- [7] 马廷泽. 长春市宽城区 2010—2011 年手足口病发病现状及经济负担研究[D]. 长春: 吉林大学, 2012.
- [8] 陈茜. 大数据对因果思维影响——从传统的因果关系向相关关系转变[J]. *现代经济信息*, 2015, 30(15): 87.
- [9] 宋辞, 裴韬. 基于特征的时间序列聚类方法研究进展[J]. *地理科学进展*, 2012, 31(10): 1307-1317.
- [10] Wang X, Smith K, Hyndman R. Characteristic-based clustering for time series data[J]. *Data Mining and Knowledge Discovery*, 2006, 13(3): 335-364.
- [11] Ooi MH, Solomon T, Podin Y, et al. Evaluation of different clinical sample types in diagnosis of human enterovirus 71-associated hand-foot-and-mouth disease[J]. *J Clin Microbiol*, 2007, 45(6): 1858-1866.
- [12] Wang JF, Guo YS, Christakos G, et al. Hand, foot and mouth disease: spatiotemporal transmission and climate[J]. *Int J Health Geogr*, 2011, 10: 25.
- [13] 肖革新, 胡跃华, 于石成, 等. 2008—2011 年中国 5 岁以下儿童手足口病流行特征分析[J]. *疾病监测*, 2012, 27(12): 932-936.
- [14] Wang X, Smith M, Miles K, Hyndman R. Rule induction for forecasting method selection: meta-learning the characteristics of univariate time series[J]. *Neurocomputing*, 2009, 72(10): 2581-2594.
- [15] Shumway R, Stoffer D. Time series analysis and its applications with R examples[M]. New York: Springer, 2009: 3-17.
- [16] 许闯. 2011—2013 年上海枫林社区手足口病流行病学特征分析[J]. *健康教育与健康促进*, 2014, 9(4): 276-279.
- [17] 徐琼, 徐莉丽, 张华一. 青海省 2011—2012 年手足口病病原学监测及流行特征分析[J]. *医学动物防制*, 2014, 35(9): 1013-1014.
- [18] 康瑞华. 2008—2013 年登陆广东、福建、海南的热带气旋对手足口病的影响[D]. 济南: 山东大学, 2016.
- [19] 罗柏生, 张莉. 2009—2012 年重庆市北碚区手足口病流行病学分析[J]. *华南预防医学*, 2014, 55(1): 51-53.

(责任编辑: 冉明会)