·新型冠状病毒肺炎疫情防控·

基于大数据与有效距离模型的突发急性传染病宏观预警及防控管理工作探讨：COVID-19疫情数据的启示

王震坤 陈知水 杜艾桦 王从义 刘虹 王子伟 胡继发
华中科技大学同济医学院附属同济医院科研处，武汉 430030
通信作者：胡继发，Email：jfahu@tjh.tjmu.edu.cn

【摘要】 目的 从宏观视角利用新型冠状病毒肺炎（COVID-19）疫情数据，分析有效距离与疫情传播轨迹、时间和规模之间的关系，为今后疫情防控提供科学依据。**方法** 收集整理截至2020年2月23日我国各地COVID-19首例确诊患者的住院治疗/隔离治疗日期以及累计确诊病例数，利用“百度迁徙”基于地理位置的服务大数据平台（LBS）获取武汉市到各地的迁出人口比例数据，建立有效距离模型和线性回归模型，从省级和市级层面分别对有效距离与疫情抵达时间及累计确诊病例级数的关系进行分析。**结果** 不论在省级层面还是市级层面上，武汉市到目的地的有效距离与COVID-19疫情抵达时间及累计确诊病例级数都存在明显的线性关联，各线性模型回归系数均有统计学意义（$P<0.001$）。在省级层面上，有效距离可以解释其与抵达时间模型71%的变异，解释其与累计确诊病例级数模型90%的变异；在市级层面上，有效距离可以解释其与抵达时间模型66%的变异，解释其与累计确诊病例级数模型85%的变异。**结论** 模型拟合程度较好，LBS大数据与有效距离模型能够用于对疫情传播轨迹、时间和规模等进行估计，为突发急性传染病宏观预警及防控管理工作提供有益参考。

【关键词】 新型冠状病毒肺炎；有效距离；人口迁徙；传染病；预警防控
基金项目：国家自然科学基金（81903396）；中国科协调研宣传部资助重点课题（20200608CG111302）
DOI：10.3760/cma.j.cn112338-20200306-00269

Discussion on early warning, prevention and control of emerging infectious diseases from a macroscopic perspective based on big data and effective distance model: enlightenment of COVID-19 epidemic data in China

Wang Zhenkun, Chen Zhishui, Du Aihua, Wang Congyi, Liu Hong, Wang Ziwei, Hu Jifa
Department of Academic Research, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China
Corresponding author: Hu Jifa, Email: jfahu@tjh.tjmu.edu.cn

【Abstract】 Objective To provide a system for warning, preventing and controlling emerging infectious diseases from a macroscopic perspective, using the COVID-19 epidemic data and effective distance model. **Methods** The dates of hospitalization/isolation treatment of the first confirmed cases of COVID-19 and the cumulative numbers of confirmed cases in different provinces in China reported as of 23 February, 2020 were collected. The Location Based Service (LBS) big data platform of “Baidu Migration” was employed to obtain the data of the proportion of the floating population from Wuhan to all parts of the country. Effective distance models and linear regression models were established to analyze the relationship between the effective distance and the arrival time of the epidemic as well as the number of cumulative confirmed cases at provincial and municipal levels. **Results** The arrival time of the epidemic and the cumulative number of confirmed cases of COVID-19 had significant linear relationship at both provincial and municipal levels in China, and the regression coefficients of each linear model were significant ($P<0.001$). At the provincial level, the effective distance could explain about 71% of the variation of the model with arrival time along with around 90% of the variation for the model in the cumulative confirmed case magnitude; at the municipal level, the effective distance could explain about 66% of the variation for the model in arrival time, and about 85% of the variation of the model with the cumulative confirmed case magnitude. **Conclusions** The fitting degree of the models are good. The LBS big data and effective distance model can be used to

estimate the track, time and extent of epidemic spread to provide useful reference for early warning, prevention and control of emerging infectious diseases.

【Key words】 COVID-19; Effective distance; Population floating; Infectious diseases; Early warning, prevention and control

Fund programs: National Natural Science Foundation of China (81903396); Key Project Supported by Research and Publicity Department of China Association for Science and Technology (20200608CG111302)

DOI:10.3760/cma.j.cn112338-20200306-00269

2019年12月以来，武汉市陆续发现多例不明原因肺炎病例，现已证实为新型冠状病毒感染引起的急性呼吸道传染病——新型冠状病毒肺炎（COVID-19）[1-2]。此次疫情传播速度快、感染范围广、防控难度大[3]。战胜疫病离不开科学支撑。如果能预估突发急性传染病在未来传播的轨迹、时间和规模，无疑会使疫情防控工作更好地做到有的放矢。Brockmann和Helbing[4]提出使用有效距离模型的概念，通过采用城市间航空客流量矩阵计算有效距离来分析预测全球传染病的蔓延轨迹和相对时间。但由于我国人口迁徙出行方式的多样化，仅依靠航空客流量无法代表国内城市间人口迁徙的流量。本研究拟将基于地理位置的服务大数据用于有效距离模型，对此次COVID-19疫情数据进行分析，探讨在病原学及流行病学等特征未知的情况下，如何根据城市间的人口相互迁徙量，研究估计疫情在城市间的传播轨迹以及相对的抵达时间和疫情规模，为今后相关疫情的预警防控等方面工作提供参考。

资料与方法

1. 数据来源：从国家卫生健康委员会及全国31省（自治区、直辖市）卫生健康委员会官方网站发布的COVID-19疫情每日报告以及各省（不包括中国香港、澳门和台湾地区数据）地方官方媒体对各自首例COVID-19相关报道中，按省份层面和城市层面整理COVID-19累计确诊病例数，时间截至2020年2月23日24：00，以及各首例COVID-19确诊患者的住院治疗/隔离治疗日期。将整理的各地首例确诊患者的住院治疗/隔离治疗日期作为抵达时间（arrival time，AT）。因不同省份间和不同城市间的COVID-19累计确诊病例数相差巨大，为方便线性回归模型的建模，本研究使用定义取累计确诊病例数（N）的对数为累计确诊病例级数（M），即：$M=\log_{10}N$。

本研究所使用的省份层面和城市层面人口迁徙数据为“百度迁徙”（http://qianxi.baidu.com）平台中统计的移动互联网大数据。该平台利用基于地理位置的服务（Location Based Service，LBS）技术，为数十万款APP提供免费、优质的定位服务，其数据来源于百度地图和第三方用户每日数十亿次的定位数据统计[5]，实时、动态、直观地展示了区域间人口日常流动，真实地记录了数以亿计的人口迁徙轨迹[6-7]。本研究利用“百度迁徙”基于地理位置的服务大数据平台，获取2020年1月10—24日武汉市到全国31省的人口迁出比例数据以及到迁出量排前100位的城市的人口迁出比例数据（即武汉市迁至各目的地的人数占武汉市迁出总人口数的比例）。

2. 有效距离模型：Brockmann和Helbing[4]在探讨影响传染病传播情况的因素时，提出了一种非传统意义上地理距离的新方法——有效距离（effective distance）模型。该模型基于两地之间人口流动（用全球航空客运量代表）构造矩阵计算有效距离，发现有效距离是传染病传播轨迹的有效衡量，并且传染病最初暴发地（outbreak location）与各地之间的有效距离与传染病到达该地的时间成固定比例。该模型已在2009年的甲型H1N1流感疫情和2003年的SARS疫情数据上得到验证[4]。

有效距离具体计算原理：用包含$0\leqslant P_{mn}\leqslant 1$元素的矩阵$P$来量化网络中直接相连的两节点中从$n$节点（最初暴发地）到$m$节点（目的地）的迁徙人口比例，则从$n$节点到$m$节点的有效距离$d_{mn}$的计算公式为：$d_{mn}=(1-\log P_{mn})\geqslant 1$。有效距离一般具有非对称性（asymmetric）特征，故有$d_{mn}\neq d_{nm}$。在有效距离的基础上，可以定义有序路径的有向长度$\lambda(\Gamma)$，其中$\Gamma=\{n_1,\cdots,n_L\}$为沿有序路径分支的有效长度之和。则在网络中任意两节点，从n节点到m节点的有效距离d_{mn}，可由n节点到m节点间最短路径的有效长度定义：$d_{mn}=\min\limits_{\Gamma}\lambda(\Gamma)$，同样的，$d_{mn}\neq d_{nm}$[4]。

3. 统计学分析：采用Python 3.6.9与Excel 2019软件对数据进行抓取及整理；采用SPSS 23软件进行有效距离模型和线性回归模型的建模计算。

结　果

1. 人口迁出的情况：2020年1月10—24日，在省级层面，武汉市迁出人口主要目的地为：湖北省（占迁出人口总量的69.40%）；河南省（占迁出人口

总量的5.68%）；湖南省（占迁出人口总量的3.48%）。在市级层面，武汉市迁出人口主要目的地为：孝感市（占迁出人口总量的13.80%）；黄冈市（占迁出人口总量的13.04%）；荆州市（占迁出人口总量的6.54%）。

2. 有效距离的计算：省级有效距离在1.35～10.21之间，市级有效距离在2.98～8.01之间。需注意的是由于“百度迁徙”中仅提供迁出量排前100位城市的人口迁徙数据，根据与地理距离对比，有效距离模型实际上是对真实地理距离存在一定差异，展示了由网络驱动的复杂传染现象背后隐藏的一种几何结构。

3. 省级层面有效距离与疫情抵达时间及累计确诊病例级数的关系：武汉市到省级目的地的有效距离与COVID-19疫情抵达时间的关系如图1A所示，散点图存在线性趋势，对其建立线性回归模型后结果表明，回归系数为3.59（95%*CI*：2.73～4.45）；经*t*检验，$t=8.50$，$P<0.001$；回归方程为AT＝13.41＋3.59D，决定系数$R^2=0.71$，表明有效距离可以解释该模型71%的变异。武汉市到省级目的地的有效距离与目的地COVID-19累计确诊病例级数的关系如图1B所示，散点图存在线性趋势，对其建立线性回归模型后结果表明，回归系数为−0.40（95%*CI*：−0.44～−0.35）；经*t*检验，$t=-16.58$，$P<0.001$；回归方程为$M=4.86-0.40$D，决定系数$R^2=0.90$，表明有效距离可以解释该模型90%的变异。

4. 市级层面有效距离与疫情抵达时间及累计确诊病例级数的关系：武汉市到市级目的地的有效距离与COVID-19疫情抵达时间的关系如图2A所示（因数据可及性，仅展示本研究能获取抵达时间的51个城市数据），散点图存在线性趋势，对其建立线性回归模型后结果表明，回归系数为4.25（95%*CI*：3.38～5.11）；经*t*检验，$t=9.82$，$P<0.001$；回归方程为AT＝6.86＋4.24D，决定系数$R^2=0.66$，表明有效距离可以解释该模型66%的变异。武汉市到市级目的地的有效距离与目的地COVID-19累计确诊病例级数的关系如图2B所示，散点图存在线性趋势，对其建立线性回归模型后结果表明，回归系数为−0.38（95%*CI*：−0.41～−0.35）；经*t*检验，$t=-23.59$，$P<0.001$；回归方程为$M=4.60-0.38$D，决定系数$R^2=0.85$，表明有效距离可以解释该模型85%的变异。

5. 不同时期省级、市级层面有效距离与累计确诊病例级数的关系：为进一步验证有效距离与累计确诊病例级数间关系的稳定性和适用范围，对武汉市到省级、市级目的地的有效距离与不同时期（武汉市关闭离汉通道后的第1、7、14、21、28天）COVID-19疫情累计确诊病例级数建立线性回归模型，结果表明各模型回归系数均具有统计学意义（$P<0.001$）。见表1。并且除市级目的地的第1天模型的决定系数较差外，其他时期的模型决定系数均较高。

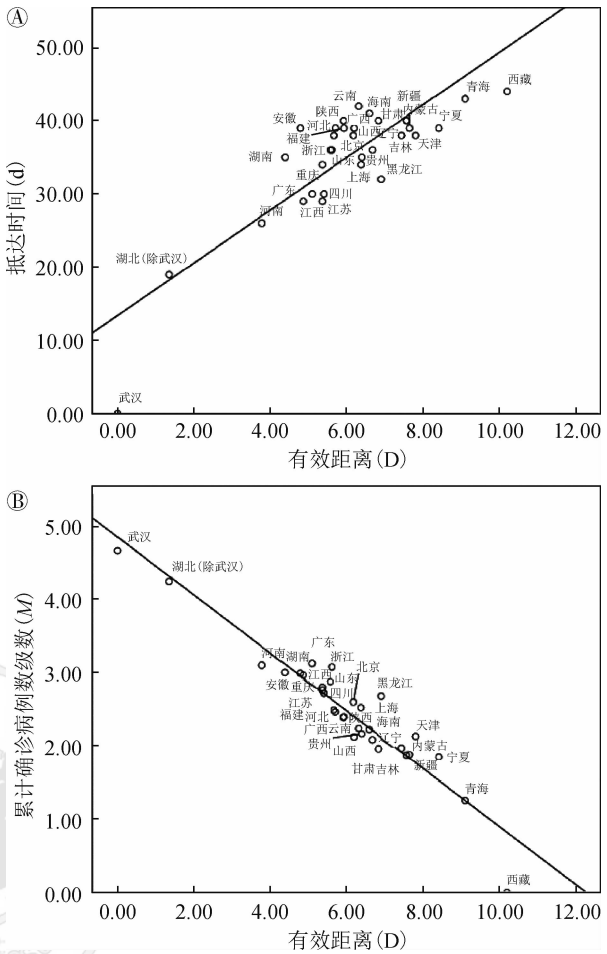

图1 武汉市到省级目的地的有效距离与COVID-19疫情抵达时间（A）及累计确诊病例级数（B）的关系

讨　论

目前国内关于突发急性传染病宏观预警及防控管理工作方面的研究较为少见。现有的相关研究对象和视角均集中在病原、患者、感染人群和密切接触人群等层面，在这些层面上疾病的传播因素都异常复杂、难以掌握，它跟病毒携带者的身体条件、所处环境的情况、气候情况、接触者的条件等都密切相关[8-10]。但是如果研究能够以更为宏观的研究视角，将一个城市抽象为一个节点（node）作为研究对象，那么完全可能对城市间的突发急性传染病传播情况做出推断。

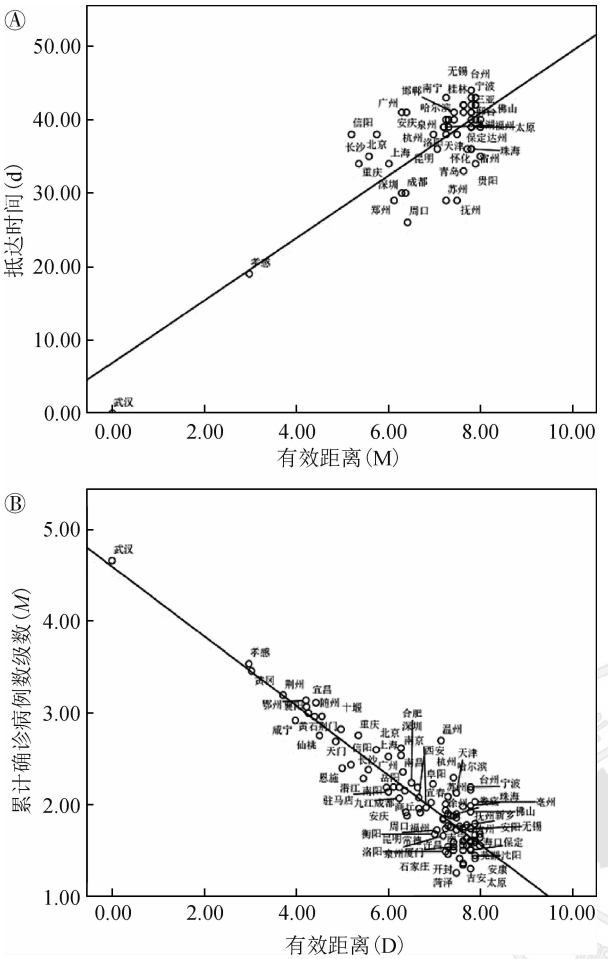

图2　武汉市到市级目的地的有效距离与COVID-19疫情抵达时间及累计确诊病例级数的关系

本研究尝试将LBS大数据用于有效距离模型，对此次COVID-19疫情数据进行分析，结果表明不论在省级层面还是市级层面上，武汉市到目的地的有效距离与COVID-19疫情抵达时间及累计确诊病例级数都存在明显的线性关系。对于单变量线性回归模型来讲，模型拟合程度较好，提示根据城市间人口迁徙数据将世界“变形”后得到的有效距离对于疫情宏观预警及防控管理工作具有一定意义。此外，对武汉市到省级、市级目的地的有效距离与不同时期COVID-19疫情累计确诊病例级数建立线性回归模型，各模型回归系数均具有统计学意义（$P<0.001$）；且除市级目的地1月24日模型的决定系数较差外（这应该与早期因检测能力或速度限制导致部分城市确诊病例为0有关），其他时期的模型决定系数均较高，说明该方法具有较好的稳定性。

当将城市抽象为节点进行聚焦的时候，会发现城市间的交通流量决定了城市间的病毒传播，而交通流量越大的两个城市在一般情况下，病毒携带者也越可能来往[4,11]。根据有效距离模型，随着交通方式的便利，决定疫情传播轨迹的不在是依据传统意义上的地理距离，而是依据疫情暴发地到不同目的地的有效距离的远近；并且即使疫情的流行病学参数未知，仍能借助有效距离来预估疫情的相对到达时间[4, 11-12]。本研究结果展示的有效距离与COVID-19疫情抵达时间之间的线性趋势亦表明，疫情的宏观轨迹大体上是依据最初暴发地到其有效距离进行由近到远的传播，这能为今后的疫情防控管理工作提供指导。例如，在1月20日晚间确认COVID-19存在人传人现象后，尽管在21日疫情通报信息中仅出现河南、湖南、广东、重庆、浙江、北京等地的确诊病例，但可推断有效距离更近的湖北省（除武汉市外）作为传播轨迹最近的地区理论上应已出现疫情，可根据地级市有效距离的远近尽快展开排查防控工作。

虽然有效距离这一单一变量能够很大程度解释疫情抵达时间的变异，但目前其只能用于对抵达时间的大体预估，因为COVID-19疫情数据表明同样有效距离的城市抵达时间尚存在一定差异。这可能与将公开报道中首例确诊患者的住院治疗/隔离治疗日期作为抵达时间的准确性限制，以及还有其他重要解释变量未纳入考虑有关。而有效距离这一单一变量能够对疫情规模（累计确诊病例级数）的绝大部分变异进行解释，并且在疫情不同时期这一方法

表1　2020年武汉市到省级、市级目的地的有效距离（D）与不同时期COVID-19疫情累计确诊病例级数（M）的线性回归模型

目的地	日期(月-日)	回归系数(95%*CI*)	回归系数t检验	P值	回归方程	决定系数(R^2)
省级	01-24	−0.31(−0.36～−0.25)	−11.36	0.000	M=3.00－0.31D	0.81
	01-31	−0.35(−0.40～−0.30)	−14.58	0.000	M = 4.12－0.35D	0.88
	02-07	−0.38(−0.43～−0.33)	−15.99	0.000	M = 4.65－0.38D	0.90
	02-14	−0.41(−0.46～−0.36)	−17.03	0.000	M = 4.94－0.41D	0.91
	02-21	−0.42(−0.47～−0.37)	−16.83	0.000	M = 5.00－0.42D	0.90
市级	01-24	−0.18(−0.25～−0.12)	−5.66	0.000	M = 1.74－0.18D	0.25
	01-31	−0.32(−0.36～−0.28)	−18.09	0.000	M = 3.73－0.32D	0.77
	02-07	−0.35(−0.38～−0.32)	−20.66	0.000	M = 4.27－0.35D	0.81
	02-14	−0.37(−0.41～−0.34)	−22.93	0.000	M = 4.54－0.37D	0.84
	02-21	−0.38(−0.41～−0.35)	−23.51	0.000	M = 4.60－0.38D	0.85

稳定性较好，提示其可以用于今后疫情规模的分析和早期预警工作。例如，前文论述到1月21日推断有效距离更近的湖北（除武汉市外）作为传播轨迹最近的地区理论上应已出现疫情，而在25日政府对24日疫情通报信息中有效距离很近的咸宁、襄阳、黄石均尚无确诊病例，但根据有效距离和疫情规模的线性关系，可以计算出24日这3个城市累计确诊病例级数在1.15～1.22之间，即可推测这3个城市24日时其累计确诊病例数在14～17例左右。

值得注意的是，尽管有效距离能够解释疫情规模的绝大部分变异，但现实情况由于其复杂性往往存在例外，比如图2A中温州市这样里拟合线偏离较远的城市需要具体分析。根据目前的信息可推测温州市的疫情规模很高的原因很可能与如下因素有关：从武汉市前往温州市的人经商比例高，经商人群相较于普通务工者每天接触的人数更多，受感染的机会更大；不会在法定节假日才回家，可能较早回到温州市，这部分人群在“百度迁徙”1月10—24日间的数据之外；更加热衷于春节前后参加一些人口密集民俗活动的习俗等。

本研究存在局限性。第一，虽然“百度迁徙”是基于数以亿计人口迁徙的LBS大数据平台，但也存在被采集群体偏向性较强等问题[6]，用其进行有效距离模型的构建不可避免地会产生一定偏差。但就目前来说，该平台提供的大数据已经是本研究能获取到的较好代表性数据。第二，由于此次COVID-19疫情的特殊性，其各地发布的首例确诊患者的确诊时间十分集中，而患者回忆的发病时间主观性较大，两者均无法用作疫情抵达时间建模，故本研究采用的是从各地地方官方网站对其首例确诊患者的报道中整理的住院时间/隔离治疗时间作为疫情抵达时间进行研究。但部分城市并未对首例确诊患者的详情进行报道，此外不可避免的是，当地第2例或第3例确诊患者的住院时间/隔离治疗时间是可能早于首例确诊患者的。今后若能获取完整全面的各地确诊患者中最早住院时间/隔离治疗时间进行建模，模型拟合度应会进一步提升。第三，虽然本研究显示采取有效距离单一变量对疫情传播轨迹、时间和规模能够进行一定程度上的估计，但也能发现疫情抵达时间部分变异只靠有效距离这一变量无法解释，这提示今后的研究采用更精准的数据进行建模后，若仍无法明显提高有效距离对模型的解释程度，则应进一步纳入更多的其他变量进行建模研究。

利益冲突 所有作者均声明不存在利益冲突

志谢 本文是作者在医院抗疫工作岗位上（医疗生活物资调配、人员物资通行入关、援汉医疗队对接等）利用夜晚休息时间完成，提笔行文间无不受白天工作所见所闻打动，在此特向所有为此次疫情爱心捐赠物资的单位和个人、向所有抗疫“逆行者”和援汉援鄂的医疗队成员以及向一直奋战在抗疫一线工作的医务工作者致以崇高敬意和衷心感谢！

参 考 文 献

[1] Zhou P, Yang XL, Wang XG, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin[J]. Nature, 2020, 579(7798): 270-273. DOI: 10.1038/s41586-020-2012-7.

[2] Lu RJ, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding[J]. Lancet, 2020, 395(10224): 565-574. DOI: 10.1016/S0140-6736(20)30251-8.

[3] 郭岩，黄旸木，黄捷，等. 新型冠状病毒肺炎疫情的全球流行现状和其对中国的影响及政策建议[J]. 中华流行病学杂志，2020, 41(5): 643-648. DOI: 10.3760/cma.j.cn112338-20200301-00222.
Guo Y, Huang YM, Huang J, et al. COVID-19 Pandemic: global epidemiological trends and China's subsequent preparedness and responses[J]. Chin J Epidemiol, 2020, 41(5): 643-648. DOI: 10.3760/cma.j.cn112338-20200301-00222.

[4] Brockmann D, Helbing D. The hidden geometry of complex, network-driven contagion phenomena[J]. Science, 2013, 342(6164): 1337-1342. DOI: 10.1126/science.1245200.

[5] 芮绍炜. 百度大数据的应用分析[J]. 企业管理，2015,(2): 114-116. DOI: 10.3969/j.issn.1003-2320.2015.02.033.
Rui SW. Application analysis of Baidu big data[J]. Enterpr Manag, 2015,(2): 114-116. DOI: 10.3969/j.issn.1003-2320.2015.02.033.

[6] 刘望保，石恩名. 基于ICT的中国城市间人口日常流动空间格局——以百度迁徙为例[J]. 地理学报，2016, 71(10): 1667-1679. DOI: 10.11821/dlxb201610001.
Liu WB, Shi EM. Spatial pattern of population daily flow among cities based on ICT: a case study of “Baidu Migration”[J]. Acta Geogr Sin, 2016, 71(10): 1667-1679. DOI: 10.11821/dlxb201610001.

[7] 蒋小荣，汪胜兰. 中国地级以上城市人口流动网络研究——基于百度迁徙大数据的分析[J]. 中国人口科学，2017(2): 35-46.
Jiang XR, Wang SL. A study on the population flow network of cities above prefecture level in China—Analysis based on Baidu migration big data[J]. Chin J Popul Sci, 2017(2): 35-46.

[8] 曲江文，聂绍发. 传染病预测预警方法的研究进展[J]. 医学与社会，2014, 27(10): 13-15. DOI: 10.13723/j.yxysh.2014.10.005.
Qu JW, Nie SF. Progress of the methods to prediction and early warning of infectious diseases[J]. Med Soc, 2014, 27(10): 13-15. DOI: 10.13723/j.yxysh.2014.10.005.

[9] 杨维中，兰亚佳，李中杰. 传染病预警研究回顾与展望[J]. 中华预防医学杂志，2014, 48(4): 244-247. DOI: 10.3760/cma.j.issn.0253-9624.2014.04.002.
Yang WZ, Lan YJ, Li ZJ. Review and prospect of early warning research on infectious diseases[J]. Chin J Prev Med, 2014, 48(4): 244-247. DOI: 10.3760/cma.j.issn.0253-9624.2014.04.002.

[10] Wang L, Li X. Spatial epidemiology of networked metapopulation: an overview[J]. Chin Sci Bull, 2014, 59(28): 3511-3522. DOI: 10.1007/s11434-014-0499-8.

[11] Brockmann D. Understanding and predicting the global spread of emergent infectious diseases[J]. Public Health Forum, 2014, 22(3): 4.e1-4.e4. DOI: 10.1016/j.phf.2014.07.001.

[12] McLean RA. Coming to an airport near you[J]. Science, 2013, 342(6164): 1330-1331. DOI: 10.1126/science.1247830.

（收稿日期：2020-03-06）
（本文编辑：万玉立）