2022 年 1 月 数理统计与管理 Jan. 2022
第 41 卷 第 1 期 Journal of Applied Statistics and Management Vol. 41 No. 1

文章编号：1002–1566(2022)01–0037 –13
DOI: 10.13860/j.cnki.sltj.20211207–002

面向 COVID–19 传播模式的多因素影响分析

谭宏卫 [1,2]

(1. 贵州财经大学数统学院，贵州 贵阳 550025；2. 贵州省大数据统计分析重点实验室，贵州 贵阳 550025)

摘要：目前，新型冠状病毒肺炎 (COVID–19) 的传播仍在持续，其传播模式以及影响传播行为的主要因素仍有待深入挖掘。鉴于此，本文从数据分析的角度，通过构造一个特殊的多源数据集 (包括 COVID–19 历史数据、气象数据、人口迁徙数据和空间地理信息数据)，以此建立多元 Poisson 回归模型 (类 Poisson 回归) 来着重分析国内疫情的病毒传播模式及其影响因素。分析结果显示，湿度、平均每日风速、每日的降雨量等气象因素与 COVID–19 的传播模式显著相关，但与每日温度变化显著不相关。除此之外，COVID–19 的传播速度及传播范围与武汉迁出目的地的人口比例、迁入武汉来源地的人口比例以及武汉与其他城市的空间距离均有一定的关联性。全文可视化及模型分析的 R 代码见：https://github.com/thwgithub/COVID-19-Rcodes。

关键词：新型冠状病毒肺炎；多因素影响分析；可视化；Poisson 回归模型

中图分类号：O212 **文献标识码**：A

Multi-factor Influence Analysis for COVID-19 Transmission Patterns

TAN Hong-Wei[1,2]

(1. School of Mathematics and Statistics, Guizhou University of Finance and Economics, Guiyang 550025, China; 2. Guizhou Key Laboratory of Big Data Statistical Analysis, Guiyang 550025, China)

Abstract: At present, COVID-19 is still spreading, and its transmission patterns and the main factors for affecting transmission behavior need to be further mined. To this end, from the perspective of data analysis, a multivariate Poisson regression model (Quasi-Poisson) based on an elaborate multi-source dataset (including COVID-19 historical data, meteorological data, population migration and spatial geographic information data), is established, whice focuses on analyzing the transmission patterns and exploring some impact factors for the COVID-19 outbreak in China. The analysis results showed that meteorological factors such as humidity, average daily wind speed, and daily rainfall are significantly correlated with the transmission patterns of COVID-19, but not with the daily temperature change. In addition, the transmission speed and transmission range of COVID-19 are related to the proportion of the population for the destination of Wuhan out-migration, the origin of in-migration Wuhun and the spatial distance between Wuhan and other cities. All of the R codes for this paper are available on the site https://github.com/thwgithub/COVID-19-Rcodes.

Key words: COVID-19; multi-factor influence analysis; visualization; Poisson regression model

收稿日期: 2020 年 7 月 28 日 **收修改稿日期**: 2020 年 9 月 12 日
基金项目: 贵州省教育厅创新群体项目 (黔教合 KY 字 [2021]015)；贵州省大数据统计分析重点实验室 (黔科合平台人才 [2019]5103 号)。

0 引言

自 2019 年 12 月，新型冠状病毒肺炎 (Corona VIr5us Disease 2019, COVID–19) 暴发以来，迅速蔓延至全国，截至目前 (2020–06–30) 全球累积确诊病例超过 1100 万例，死亡人数超过 50 万人 (国内：确诊 85232 例，死亡 4648 人)[1]。这个突发的公共卫生事件受到全球的生物学家，流行病学家，医学家，数学家等广泛的关注，通过不同领域的共同研究，希望能对 COVID–19 的防预和控制起到实质性的作用。同时，这也是本文研究的出发点。

目前，针对 COVID–19 的多方向研究中，数理研究是最活跃的研究方向之一。数理研究主要是以数学和数理统计学作为基础工具，构建数学模型和统计模型对病毒时变传播动力、病毒传播风险、死亡风险估计、病毒增长率估计以及病毒感染病例的峰值预测等多角度研究。这些研究对探究病毒的传播规律有一定的导向作用。

数学模型模拟研究。鉴于 COVID–19 数据的有限性及案例信息的不完整性，通过建立数学模型，对 COVID–19 的暴发轨迹、防控的有效性、传染趋势等方面的模拟研究具有重要的实际意义。在 COVID–19 的暴发之初，通过随机模拟病毒传播轨迹发现，早期病毒的人传人方式 [2]。在病毒感染和流行趋势方面，一些学者用 SIR 模型以及变式 SEIR 模型来模拟，并获得一些建设性的结论 [3]。特别地，通过 SEIR 模型预测 COVID–19 感染的国内确诊病例总数在 $[70000, 90000]$，从目前的趋势看，这个模型是成功的 [3]。还有一些研究者，通过武汉的暴发案例数据研究 COVID–19 的传播动力 [4]。在防控有效性模拟研究方面，利用修正的 SEIR 模型来论证隔离防控的有效性以及模拟中国防控措施的作用 [5]。除此之外，也有研究者针对 “封城”、交通管制以及区域隔离的有效性进行模拟研究 [6]，模拟结果显示，对于 COVID–19 的高传播风险，这些应急措施是必要且有效的。对于病毒传播风险及增长率，部分研究者从微分方程的角度对这个问题展开研究 [7]，并取得一些有价值的结论。

统计模型研究。统计模拟研究是通过建立统计模型，从多个视角对 COVID–19 展开研究。假定病毒总感染数服从一个对数正态分布 [8] 的条件下，构建一个右截断的双区间删失似然模型来探究 COVID–19 的潜伏期特征 [9]，将极大地降低病毒的再传播风险。对现有的 COVID–19 数据进行模型拟合分析，也是这个领域研究的热点之一，如死亡风险估计、感染风险估计、大范围暴发风险估计及预测感染总人数等。特别地，从多源数据分析的视角，如新闻媒体报道数据、社区体检数据、交通与人流量数据来估计病毒暴发的风险，通过深层次的分析，获取了有用的信息 [10]。在感染病例总数的预测中，Logistic 回归模型 [11] 对国内 COVID–19 感染病例总数的预测结果较符合实际情况 [12]。此外，研究者们还根据一些临床数据来分析病毒的潜伏期分布及临床特征 [13]，以及利用大数据与人工智能来分析 COVID–19 的传播趋势和传播轨迹 [14]。进一步研究发现，COVID–19 的传播模式与气候条件具有一定的统计相关性；研究者们收集全球范围内的 COVID–19 数据以及相应的地区气候数据，分析得出：平均光照时间较大的区域病毒增长较快，区域平均温度与区域病例数呈中度相关 (Pearson 相关性)[15]。尽管如此，随着样本量 (确诊病例数) 的增加，这一结论有待进一步考证。

根据对上述文献的调查及分析，针对 COVID–19 多因素影响的分析模型有待进一步开发。本文利用日常报告的 COVID–19 历史数据、人口迁徙数据、气象数据和空间地理信息数据，构建一个多元的 Poisson 回归模型，对国内的 COVID–19 病毒传播模式展开研究，充分地挖掘这些数据所隐含的信息，深入揭示 COVID–19 的传播模式。

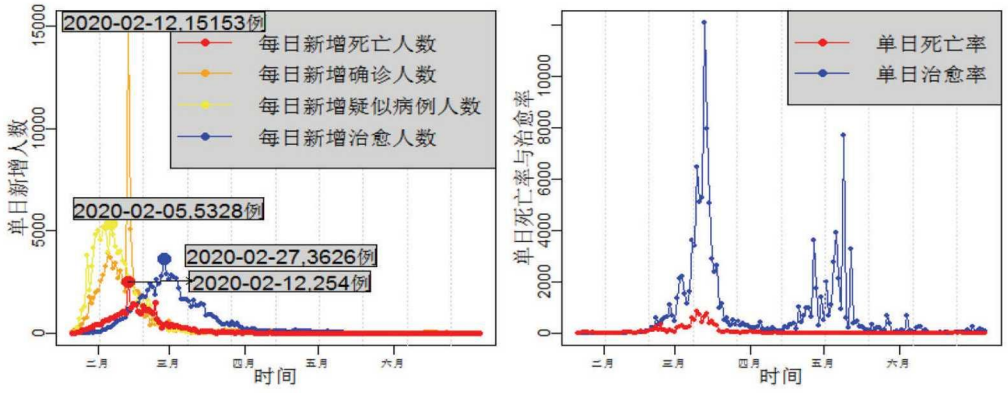
1 单源数据分析

本节内容主要对 COVID–19 的历史数据以及疫情期间 (2020–01–01 至 2020–01–23) 的

人口迁徙数据进行初步分析，前者的来源是 R 语言中的 nCOV2019 程序包 (https://github.com/GuangchuangYu/nCov2019)，这个包即能调取实时疫情数据，也能提取国内外疫情历史数据，本文着重分析国内疫情数据；后者的来源是百度迁徙大数据平台 (https://qianxi.baidu.com/2020/)。对于 COVID–19 的历史数据，主要进行一些初步的描述性分析；而对于人口迁徙数据主要研究人口迁徙强度与累积确诊病例之间的相关性。

1.1 COVID–19 历史数据可视化分析

截至目前 (2020–06–30)，全国 COVID–19 病毒累计确诊 85232 例，累计治愈 80068 例，累计死亡 4648 例，现有确诊 516 例，疑似病例 8 例，累计境外输入病例 1918 例。从这些基本数据可看出，国内疫情趋于消退态势。接下来，利用国内的每日新增数据 (2020–01–20 至 2020–06–30) 绘制曲线图及地图热图来全面地、直观地了解国内疫情传播趋势。



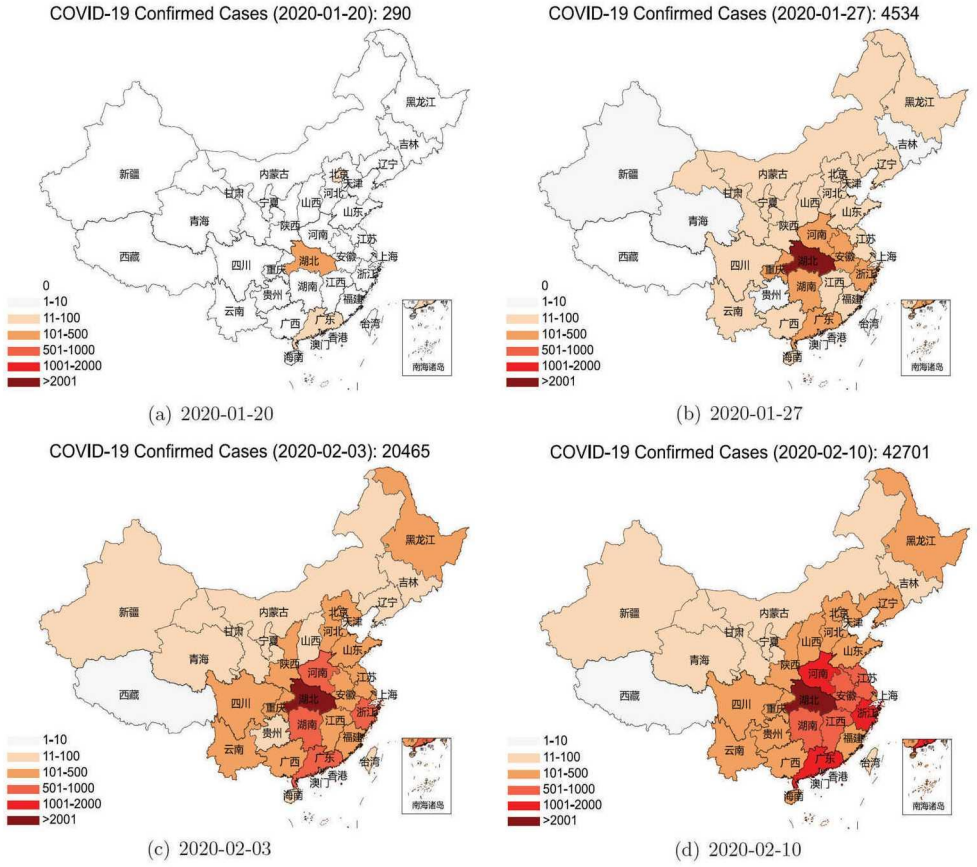
(a) 单日疫情新增人数　　(b) 单日死亡率与治愈率

图 1 国内单日疫情数据新增情况

在图 1 中，用预警系统中的级别颜色红、橙、黄、蓝分别标注死亡人数或死亡率、确诊病例数、疑似病例数及治愈病例数或治愈率。从单日新增的疫情数据图 1 (a)，可清晰地看出，二月份是国内疫情的高峰期。其中，全国单日最高新增死亡人数 254 例 (2020–02–12)，单日最高新增确诊病例 15153 例 (2020–02–12)，单日最高新增疑似病例 5328 例 (2020–02–05)，单日最高治愈病例 3626 例 (2020–02–07)。图 1 (a) 还显示，2020 年 1 月中旬至 2020 年 2 月上旬的疫情形势极不乐观，虽然在这个时段的每日治愈人数 (蓝色) 呈上升趋势，但相比每日新增确诊病例 (橙色) 及疑似病例数 (黄色) 的增长幅度，医疗系统仍承载巨大压力。注：为了清楚地展示每日死亡人数的变化趋势，图 1 (a) 的死亡人数提高 10 倍于真实情况。除此之外，从图 1 (a) 还能观察到，2020–02–05 是单日新增疑似病例最高日，其后迅速下降；而一周后 (2020–02–12) 单日新增确诊病例达到顶峰 (15153 例)，这说明：(1) 这个时期疑似病例的阳性概率极高；(2) 政府及医疗机构迅速反应，加大疑似病例检测力度及范围，做到应检尽检，应收尽收，应治尽治。单日最高确诊人数半个月后 (2020–02–12 至 2020–02–27)，单日最高治愈人数达到最大值，这说明经过医护人员的不断努力，病毒治疗效果显著改善。同样，在 2020–02–12 后，死亡人数得到有效控制。至三月中旬左右，各项数据明显下降，2020–04–14 日单日新增死亡人数彻底清零，单日疑似病例趋于个位数，单日新增确诊病例数控制在两位数以内。国内疫情形势逐步稳定，而后小范围 (黑龙江、吉林和北京) 的疫情暴发，并没有阻碍疫情向好发展趋势。图 1 (b) 的治愈率曲线 (蓝色) 出现两次高峰，分别三月份的武汉疫情及五月份黑龙江、吉林及境外输

入病例，同时还包括最近北京的疫情也得到了有效的控制。注：图 1 (b) 中的治愈率是单日治愈人数与单日新增人数的比值，同样死亡率是单日死亡人数与单日新增人数的比值，因此纵坐标治愈率出现了大于 100% 的情况。

为了更直观地了解疫情的传播速度及范围，图 2 展示了三个周的 COVID–19 病毒的传播速度。图 2 (a) 是 2020–01–20 全国疫情情况，当时全国总共确诊 290 例，仅在武汉、广东及北京出现。一个周后 (2020–01–27)，病毒迅速蔓延到全国各地 (共确诊 4534 例)，从地图上看 (图 2 (b))，只有西藏未出现病例。再过一周 (图 2 (c)，2020–02–03)，全国共确诊 20465 例几乎是上个周 (2020–01–27) 确诊病例的五倍。最后，在 2020–02–10 (图 2 (d))，全国疫情形势已经相当严峻。从图 2 可以看出，COVID–19 病毒的传播速度极快，传播范围极广。

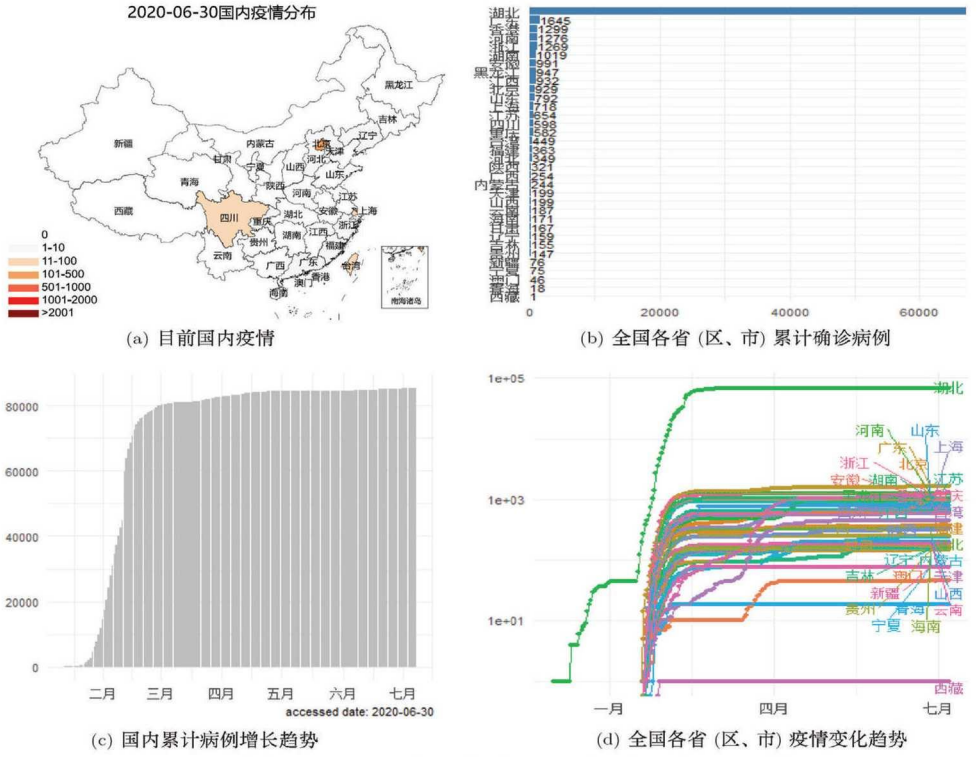


(a) 2020-01-20 (b) 2020-01-27

(c) 2020-02-03 (d) 2020-02-10

图 2 国内疫情三个周 (2020–01–20 至 2020–02–10) 的 COVID–19 病毒传播速度

接下来，通过可视化方法来观察当前 (2020–06–30) 全国范围内的疫情趋势。从图 3 (a) 可以看出，目前全国大部分地区的确诊病例数已经清零，只有北京 (324 例)，四川 (12 例)，上海 (29 例) 等少部分地区还有部分确诊病例，除北京之外的地区大部分病例是境外输入病例，这再次说明国内疫情以基本稳定。图 3 (b) 是全国 34 个省级行政区的累积病例数，除湖北以外，广东 (1645 例)、香港 (1299 例)、河南 (1276 例) 的累积病例数居前三位，而澳门 (46 例)，清海 (18 例)，西藏 (1 例) 最少。图 3 (c) 是国内累积确诊病例的历史增长条形图，可清晰看出，二月份是病例增长高峰期，到三月份以后才逐步放缓。图 3 (d) 是全国 34 个省市的疫情变化曲线，到四月中旬全国疫情才得到全面控制。即使后来小范围的疫情暴发 (黑龙江、吉林和北京)，都没有影响国内疫情向好发展的总体趋势。



(a) 目前国内疫情　(b) 全国各省 (区、市) 累计确诊病例

(c) 国内累计病例增长趋势　(d) 全国各省 (区、市) 疫情变化趋势

图 3 当前 (2020-06-30) 的疫情趋势以及各省市的变化情况

1.2 人口迁徙数据分析

为了探究国内 COVID–19 病毒的传播范围与人口迁徙强度的关系，在此以武汉的迁徙数据为例来分析。数据选取范围是 2020–01–01 至 2020–01–23 (武汉 “封城” 日)，数据主要包括两种：(1) 从武汉迁出的目的地的人口占比；(2) 迁入武汉的来源地的人口占比。武汉是全国交通流量最大的枢纽中心之一，九州通衢之地，北上河南北京天津，南下湖南贵州广东。因此，武汉的人口流动性大，病毒传播防预困难。

表 1 2020–01–01 至 2020–01–23 期间武汉与省外平均人口迁徙强度 (%) 前五省份以及与省内平均人口迁徙强度前五城市，以及相应的平均迁出率 (%)、平均迁入率 (%) 和 14 天的累积确诊病例 (人数)

省份及城市	河南	广东	湖南	江苏	浙江	孝感	黄冈	荆州	威宁	鄂州
平均迁出率	5.03	1.83	3.22	2.18	1.03	12.92	12.20	6.09	5.14	4.48
平均迁入率	3.76	6.27	2.87	2.75	2.85	12.65	10.46	4.89	5.47	5.85
平均人口迁徙强度	8.79	8.10	6.09	4.93	3.88	25.57	22.67	10.98	10.61	10.33
累积确诊病例	914	1081	772	408	1006	2141	1897	885	443	471

现考虑武汉从 2020–01–01 至 2020–01–23 期间的人口迁徙强度与病毒传播范围之间的相关性。首先，定义某一时刻某地的人口迁徙强度。设在 t 时刻，从某地 P_1 迁出到某一目的地 P_2 的人口，占 P_1 地总迁出人口比例为 Q_{out} (迁出率)，同时刻从目的地 P_2 迁入 P_1 地的人口，占 P_1 地总迁入人口比例为 Q_{in} (迁入率)，则定义 t 时刻 P_1 地的人口迁徙强度 Q 为：

$$Q = Q_{in} + Q_{out}.$$

由上式定义可得，某一时段的平均人口迁徙强度为 $\overline{Q} = \overline{Q}_{in} + \overline{Q}_{out}$，其中 $\overline{Q}_{in}$，$\overline{Q}_{out}$ 分别表示平均迁入率和平均迁出率。若 Q 或 $\overline{Q}$ 越大，则表明 P_1 地与 P_2 地之间的人口流动性越大；反之，则越小。表 1 是武汉与省外平均人口流动性前五个省级行政区和前五个省内城市，同时还包括平均迁出率，平均迁入率及 14 天后的累积确诊病例数。可以看出，14 天后的累积确诊病例与 2020–01–01 至 2020–01–23 期间武汉与全国其他城市之间的平均人口迁徙强度呈明显的正相关趋势。

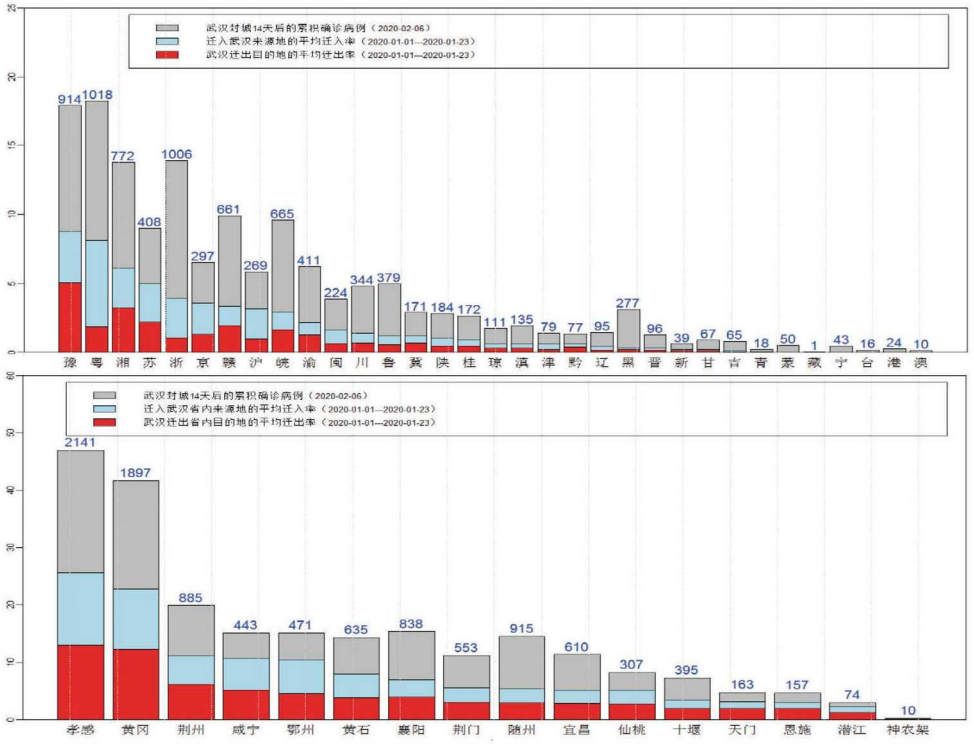


图 4 武汉在 2020–01–01 至 2020–01–23 (武汉 “封城” 日) 期间的平均人口迁徙强度 (平均迁出率 (红色)+ 平均迁入率 (浅蓝色)) 与 14 天后全国各省市的累积确诊病例 (灰色)(上半部分是除湖北之外的 33 个省级行政区，下半部分是除武汉外的湖北省内城市)

表 2 Pearson 相关系数与距离相关系数 (dcor)

	$\overline{Q}$ 与 Confirm14		$\overline{Q}_{in}$ 与 Confirm14		$\overline{Q}_{out}$ 与 Confirm14	
	省外	省内	省外	省内	省外	省内
Pearson 相关系数	0.89	0.93	0.86	0.91	0.80	0.95
dcor	0.89	0.90	0.87	0.87	87	0.92

为了更全面地展示人口迁徙强度与累积确诊病例之间的关系，将武汉在 2020–01–01 至 2020–01–23 期间的平均人口迁徙强度 (平均迁出率 (红色)+ 平均迁入率 (浅蓝色)) 以及 14 天后 (2020–02–06) 全国各省市累积确诊病例 (浅蓝色) 的总体趋势展示在图 4 中，上半部分是除湖北之外的 33 个省级行政区，下半部分是除武汉外的湖北省内城市，其中累积病例数坐标尺度缩小 100 倍。图 4 上半部分显示，这期间与武汉平均人口迁徙强度最大的省份是河南 (8.79%)，其后是广东 (8.10%)、湖南 (6.09%)，其中有 14 个省级行政区的平均人口迁徙强度在 1% 以上。其次，观察 “封城”14 天后 (2020–02–06) 的各省市累积病例，广东 (1018 例)、浙江

(1006 例) 和河南 (914 例) 最高。而与武汉平均人口迁徙强度较低的省份，其累积确诊病例也相对较低，如西藏、新疆、青海等地区。这再一次说明，平均人口迁徙强度与 14 天后的累积确诊病例数呈明显的正相关趋势。同样，从湖北省内人口迁徙情况也能得出类似的结论 (图 4 下半部分)，其中孝感和黄冈的平均人口迁徙强度 (25.57%，22.67%) 相对较高，相应的累积病例数也较高 (孝感：2141 例，黄冈：1897 例)，较低的是潜江 (2.14%，74 例) 和神农架 (0.06%，10 例)。

为了从统计的角度进一步来验证这种相关性，分别计算武汉在 2020–01–01 至 2020–01–23 期间的平均人口强度、平均迁出率及平均迁入率与 14 天后 (2020–02–06) 全国各省市累积确诊病例之间的 Pearson 相关系数以及距离相关系数 [16] (Distance Correlation，dcor)，前者度量变量之间的线性相关性，后者能度量变量之间所有类型的相关性，包括线性和非线性 (表 2)。便于叙述，用 Confirm14 表示 14 天后的累积确诊病例。表 2 表明，武汉人口迁徙强度与 Confirm14 具有较强的统计相关性。同时，这也充分地证实武汉 “封城” 措施的必要性及合理性。

2 多因素影响分析

为了深入探究国内 COVID–19 病毒的传播模式，本节将利用 COVID–19 的历史数据、人口迁徙数据、气象数据以及空间地理信息数据构造一个特殊的数据集，并对此建立多元 Poisson 回归模型进行多因素影响分析。首先，简要介绍 Poisson 回归模型；其次，介绍数据集的构成及其来源；最后，建立多元的 Poisson 回归模型，并分析病毒的传播行为及其影响因素。

2.1 Possion 回归模型简介

Possion 回归模型是一种典型的广义线性回归模型 [17]，这个模型对一些随机计数变量具有较好的拟合效果。设因变量 $Y \sim Po(\mu)$，其中 $Po(\mu)$ 表示参数为 μ 的 Possion 分布，则 $E(Y)=\mu$。在实际中，假设 $\mu = n_i\theta_i$，θ_i 为影响因素，n_i 为对应的影响系数，$i=1,2,\cdots,n$，对于 θ_i 常用如下的模型拟合：

$$\theta_i = \exp(\boldsymbol{x}_i\boldsymbol{\beta}),$$

则 Possion 回归模型为

$$E(Y)=\mu = n_i\exp(\boldsymbol{x}_i\boldsymbol{\beta}), \qquad Y\sim Po(\mu),$$

其中，$\boldsymbol{x}_i$ 是自变量，$\boldsymbol{\beta}$ 为待估参数。由上式易知，Possion 回归模型的联接函数 (Link Function) 为对数函数，于是上式模型可转化为如下形式：

$$\log\mu = \log n_i + \boldsymbol{x}_i\boldsymbol{\beta},$$

上式模型即为 Possion 回归模型的常用形式，$\log n_i$ 为常数项或偏置项，这个模型的参数 $\boldsymbol{\beta}$ 常利用极大似然法来估计。

2.2 多源数据集的构成及其来源

为了从数据分析的角度，深入探究 COVID–19 病毒的传播模式，本文选取 COVID–19 历史数据、人口迁徙数据、气象数据以及空间地理信息数据来全面地分析这个问题。由于在 2020–01–23 日武汉 “封城”，之后全国大部分城市均处于封闭状态，所以人口迁徙数据选取

2020–01–01 至 2020–01–23 期间的数据，而迁徙的城市主要考虑从武汉迁出的目的地以及迁入武汉的来源地城市。根据流行病学原理，一个典型病毒患者在迁徙的过程中感染另一个人，这个人一般不会在当日有明显的感染症状 (尤其是无症状感染者)，可能要经过一个感染周期甚至多个周期才会有典型症状，而后检测确诊。因此，本文只考虑与人口迁徙发生日对应的 14 天后的新增确诊病例数以及相应的累积确诊病例数和累积治愈病例数等历史数据，如 2020–01–01 的迁徙数据对应 2020–01–15 日的新增病例数，以此类推。因此，COVID–19 历史数据的选取范围是 2020–01–15 至 2020–02–06，即当日的人口迁徙对 14 天后的新增确诊病例产生影响。事实上，表 1、2 的结果已经验证了这个观点的正确性。与 2020–01–15 至 2020–02–06 期间的 COVID–19 历史数据相对应，气象数据也选取这个期间的数据。对于空间地理信息数据，本文选取武汉与全国 320 个城市的空间距离数据。这个数据集一共包括 31 个省级行政区 (港澳台除外)，320 个地级市，3859 个样品和 16 个变量 (在建模过程中逐一引入相应的变量)。值得强调的是，由于大部分城市统计病例数据的初始时间是 2020–01–23 左右，因此样本容量只有 3859 个。

数据来源。COVID–19 历史数据同样从 R 语言中的 nCOV2019 程序包调取；人口迁徙数据同样来自百度迁徙大数据平台。气象数据来自后知气象大数据平台 (http://hz.zc12369.com/home)；而武汉与全国 320 个城市的空间距离数据来自高德地图 (https://www.amap.com/)，使用地图中的测距功能获得。

2.3 多源数据融合的 Possion 回归模型

在本节内容中，通过建立多源数据融合的 Possion 回归模型来深入分析 COVID–19 传播模式及其非病理学影响因素。与上一节单独分析 COVID–19 历史数据和人口迁徙数据一样，在此首先单独对气象数据建模；然后，建立多源数据融合的 Possion 回归模型。在没有特别说明的情况下，模型的因变量均选择全国 320 个城市在病例数据统计开始日至 2020–02–06 期间的每日新增确诊病例数，统一用 $newconfirm$ 表示。

表 3 每日新增病例与每日气象数据的 Poisson 回归模型的参数估计及其参数检验

Coefficients	Estimate	Std.Error	z_value	$\Pr(>\|z\|)$
Intercept	−1.453 e+01	2.930 e−01	−49.595	< 2 e−16 ***
$minT$	−4.477 e−02	2.192 e−03	−20.421	< 2 e−16 ***
$maxT$	8.154 e−02	2.269 e−03	35.941	< 2 e−16 ***
H	9.501 e−03	7.606 e−04	12.491	< 2 e−16 ***
$windv$	−2.412 e−02	1.646 e−02	−1.465	0.143
$windc$	3.092 e−02	2.579 e−02	1.199	0.230
P	1.507 e−02	2.756 e−04	54.703	< 2 e−16 ***
$visibility$	−1.979 e−02	1.211 e−03	−16.337	< 2 e−16 ***
$rain$	−1.905 e−02	1.432 e−03	−13.302	< 2 e−16 ***

注:“*” 表示参数显著性程度，即非常显著 “***”，比较显著 “**”，显著 “*”，无 “*” 号则表示不显著。下表同。

2.3.1 气候变化对 COVID–19 传播的影响

目前，气候因素对 COVID–19 的传播模式是否有影响，仍无定论。本部分内容利用气象数据对 COVID–19 的每日新增确诊病例建立 Poisson 回归模型，其中气象数据中的 8 个指标作为自变量 (注：气象数据共 9 个指标，为保证模型自变量之间的独立性，每日平均温度 $aveT$ 不进

入模型)，分别是每日最高气温 $maxT$ (°C)、每日最低气温 $minT$ (°C)、每日湿度 H (%)、每日平均风速 $windv$ (m/s)、每日平均风级 $windc$、大气压 P (hpa)、能见度 $visibility$ (km) 和总降水量 $rain$ (mm)。表 3 是这个 Poisson 回归模型的参数估计及其参数检验，可以看出，2020–01–15 至 2020–02–06 期间的每日新增确诊病例 $newconfirm$ 与当日的最低气温 $minT$、最高气温 $maxT$ 及降水量 $rain$ 等气候因素显著相关，而与每日的平均风速 $windv$ 和风级 $windc$ 显著不相关。

表 4 每日新增病例与每日天气数据的类 Poisson 回归模型的参数估计及其参数检验

Coefficients	Estimate	Std.Error	z_value	Pr(> |z|)
Intercept	−14.530584	1.875054	−7.749	1.17 e−14 ***
$minT$	−0.044771	0.014031	−3.191	0.00143 **
$maxT$	0.081538	0.014519	5.616	2.09 e−08 ***
H	0.009501	0.004868	1.952	0.05103
$windv$	−0.024115	0.105370	−0.229	0.81899
$windc$	0.030924	0.165036	0.187	0.85137
P	0.015075	0.001764	8.548	< 2 e−16 ***
$visibility$	−0.019789	0.007752	−2.553	0.01073 *
$rain$	−0.019046	0.009163	−2.078	0.03773 *

表 5 约简的类 Poisson 回归模型的参数估计及其参数检验

Coefficients	Estimate	Std.Error	z_value	Pr(> |z|)
Intercept	−13.664716	1.796634	−7.606	3.54 e−14 ***
$minT$	−0.037652	0.010620	−3.545	0.000397 ***
$maxT$	0.077639	0.012089	6.422	1.51 e−10 ***
P	0.014955	0.001756	8.518	< 2 e−16 ***
$visibility$	−0.025654	0.006901	−3.718	0.000204 ***

表 6 约简模型与全模型的 Wald 检验

模型		残差	自由度	Chisq	Pr(> $Chisq$)
Model 0: newconfirm	All Variables	3854			
Model 1: newconfirm	$maxT + aveT + P + visibility$	3850	4	6.3855	0.1721

为了验证上述模型的合理性，还需检验数据的过度离势性，即检验模型估计的均值和方差相差是否过大 (理论 Poisson 分布的均值和方差相等)。在 R 语言的 qcc 程序包中，利用 qcc.overdispersion.test( ) 函数可以执行这个检验，检验的 p 值为 0，这说明每日新增病例数据存在过度离势现象，同时也说明表 3 对应的 Poisson 模型存在不合理性。为解决这个问题，本文采用类 Poisson 回归模型 (Quasi–Poisson) 重新对气象数据建模，其参数估计结果显示在表 4 中。可以看出，表 3 (Poisson 回归) 与表 4 (类 Poisson 回归) 的参数估计是一样的，但是参数的检验原理有一定差别，前者是偏 Wald 检验，后者是 t 检验。这也导致两个模型参数检验结果不一致，如在 Poisson 回归中每日湿度 H、能见度 $visibility$ 和总降雨量 $rain$ 显著相关，而在类 Poisson 回归中三个变量是不相关和弱相关。由于数据存在过度离势现象，因此类 Poisson 模型 (表 4) 更合理。在这个模型的基础上，利用前向变量选择法获取约简模型的参数估计及其参数检验 (表 5)，其结果显示，每日新增确诊病例与当日的最高温度 $maxT$、最低温度 $minT$、大气压 P 及能见度 $visibility$ 显著相关，其对应的标准差误差 (Std.Error) 也较小，这表明约简模型拟合数据是合理的。表 6 是约简模型与全模型的 Wald 检验，其 p 值为 0.1721，这表明

两个模型对数据的拟合程度无显著差异。因此，约简模型是对气象数据建模的终模型，由表 5 的结果 (保留小数点后三位) 得如下模型：

$$\log E(newconfirm) = -13.665 + 0.078maxT - 0.038minT + 0.015P - 0.026visibility.$$

上式模型表明，在其他变量保持不变的前提下，全国 320 城市的平均温度上升 1°C，则全国的平均每日新增确诊病例数是现平均每日新增病例数的 1.08 倍；在其他变量保持不变的前提下，最低温度上升 1°C，则全国的平均每日新增确诊病例数是现有平均每日新增病例数的 0.97 倍。同理，在其他变量保持不变的前提下，平均大气压 P 增加一个单位，则全国的平均每日新增确诊病例数将是现有平均每日新增病例数的 1.06 倍。值得注意的是，在其他变量保持不变的前提下，能见度 $visibility$ 增加一个单位，则全国的平均每日新增确诊病例数是现平均每日新增病例数的 0.97 倍，即空气质量越好，能见度越高，平均每日新增病例数下降。结合表 5,6，上式模型对气象数据的拟合程度较好。尽管如此，这个模型似乎并不符合客观事实，尤其是每日新增确诊病例 $newconfirm$ 与风速 $windv$ 和风级 $windc$ 显著无关，这违背了通风有利于防疫的常识；其次，这个模型显示，新增确诊数据与温度显著相关，从目前的形势看，温差对疫情的扩散似乎没有太大的影响。因此，单独分析气象因素对 COVID–19 病毒传播模式的影响存在一定的局限性。

2.3.2 多源数据融合的 Poisson 回归模型

现考虑，建立多源数据融合的 Poisson 回归模型，自变量有 14 个，包括 COVID–19 历史数据中的 3 个变量 (每日累积病例 $cum\_confirm$、每日死亡病例 cum_dead 和每日治愈病例 cum_heal，注：疑似病例在早期 (2020–01–15 至 2020–02–06) 统计较少，很多城市都为 0，因此取消这个变量进入模型)，气象数据 8 个变量 (见表 3)，人口迁徙数据 2 个变量 (武汉迁出目的地的人口占比 wu_out 和迁入武汉来源地的人口占比 in_wu)，空间地理信息数据 1 个变量 (武汉与其他城市之间的空间距离 $wutod$)。表 7 是这个模型的参数估计及其参数检验，结果显示，每日新增病例 $newconfirm$ 与累积病例 $cum\_confirm$、累积死亡病例 cum_dead、湿度 H、平均风速 $windv$、武汉与其他城市之间的空间距离 $wutod$ 等变量显著相关，而与累积治愈病例 cum_heal、每日最低温度 $minT$、每日最高温度 $maxT$ 显著不相关。现利用前向变量选择法建立这个模型的约简模型作进一步分析。

表 8 是多源数据融合的约简类 Poisson 回归模型的参数估计及其参数检验，表 9 是约简模型 (表 8) 与全模型 (表 7) 的 Wald 检验，检验结果显示，两种模型并无显著性差异 (p 值为 0.4562)。相比表 7，表 8 中的约简模型删除了三个无显著相关的变量 (cum_heal, $minT$, $maxT$)，每日新增确诊病例与这个三个变量显著无关比较符合客观事实。对于低度显著变量能见度 $visibility$ (表 7)，加入这个变量与删除这个变量的模型，Wald 检验结果并无大的差异。因此，从模型维度上考虑，删除这个变量。根据表 8 可得如下 Possion 回归模型 (保留估计参数小数点后三位)：

$$\begin{aligned}\log E(newconfirm) = \ & -2.505 + 0.003cum\_confirm - 0.166cum\_dead - 0.286windv \\ & -0.013H + 0.375windc + 0.006P - 0.023rain + 0.079wu\_out \\ & +0.093in\_wu - 0.002wutod.\end{aligned}$$

上式模型表明，平均每日新增确诊病例与累积确诊病例 $cum\_confirm$ 正相关，与累积死亡人数 cum_dead 负相关，前者是显然的结论，而后者可解释为：死亡人数每增加一人，政府防控

措施越严格，新增确诊病例得到有效控制。在气象数据中，每日湿度 H 与每日总降雨量 $rain$ 是两个相关量，两者与平均每日新增确诊病例都呈负相关，单纯地可理解为低湿度天气更有利于病毒传播；对于平均风速 ($windv$)，在其他变量保持不变的条件下，若平均风速增加一个单位，则平均每日新增确诊病例将是现有平均每日新增确诊病例的 0.75 倍，这表明每日平均风速越高 (局部区域的通风条件越好)，则每日新增确诊病例越低，这个结论与通风有利于抑制疫情蔓延的防疫措施相吻合。风级 ($windc$) 与平均每日新增确诊病例呈正相关，这个结论与平均风速的负相关性并不矛盾，每日风级越高，并不代表每日平均风速越高，同时在构造的数据集中平均风级只有 1.52 级。在所有气象指标中，大气压 P 对每日新增确诊病例的影响最小 (影响因子为 0.006)。对于人口迁徙指标，在其他变量保持不变的条件下，武汉迁出人口占比 wu_out (%) 每增加一个单位 (一个百分点)，则迁出目的地城市的平均每日新增确诊病例约上升到原有水平的 1.08 倍；同样，在其他变量保持不变的条件下，迁入武汉的来源地人口占比 in_wu 每增加一个单位 (一个百分点)，则来源地城市的平均每日新增确诊病例约上升到原有水平的 1.10 倍。上式模型对人口迁徙数据的拟合结果说明，从两个迁徙指标对平均每日新增病例的影响来看，去过武汉的人 (1.10) 比从武汉出来的人 (1.08) 对疫情的蔓延影响稍大一点。这个结论再一次充分地证实，人口流动 (包括迁入和迁出) 对 COVID–19 的传播有较大地影响，人口流动性越大，病毒传播的速度越快，范围越广。对于空间距离指标 $wutod$，武汉与其他城市的距离每增加一个单位 (km)，则平均每日新增确诊病例数下降到原有水平的 0.99 倍，在病毒暴发的初期，这个结论比较符合客观事实。综上分析得出，上式模型较为合理地刻画了 COVID–19 病毒传播的实际情况。基于对现有数据的分析，COVID–19 的传播模式与部分气候特征、人口迁徙强度以及武汉与全国其他城市之间的空间距离具有一定的关联性。

表 7 多源数据融合的类 Poisson 回归模型的参数估计及其参数检验

Coefficients	Estimate	Std.Error	z_value	Pr(> |z|)
Intercept	−2.071 e+00	8.041 e−01	−2.575	0.010048 *
$cum\_confirm$	3.110 e−03	1.528 e−04	20.355	<2 e−16 ***
cum_heal	−4.195 e−03	2.911 e−03	−1.441	0.149669
cum_dead	−1.609 e−01	1.214 e−02	−13.253	<2 e−16 ***
$minT$	7.660 e−03	7.289 e−03	1.051	0.293347
$maxT$	8.234 e−03	7.476 e−03	1.101	0.270758
H	−1.486 e−02	2.334 e−03	−6.366	2.16 e−10 ***
$windv$	−2.588 e−01	5.434 e−02	−4.762	1.99 e−06 ***
$windc$	3.662 e−01	8.382 e−02	4.369	1.28 e−05 ***
P	5.286 e−03	7.561 e−04	6.991	3.21 e−12 ***
$visibility$	−1.166 e−02	4.590 e−03	−2.540	0.011127 *
$rain$	−2.040 e−02	5.313 e−03	−3.839	0.000125 ***
wu_out	6.893 e−02	2.707 e−02	2.546	0.010921 *
in_wu	1.128 e−01	2.758 e−02	4.089	4.43 e−05 ***
wutod	−1.528 e−03	9.716 e−05	−15.726	<2 e−16 ***

3 结语

本文主要从数据分析的角度， 利用描述性分析和模型分析相结合的分析方法对国内 COVID–19 病毒的传播模式展开研究。通过可视化分析得出，2020 年 1 月中下旬至 2020 年 2 月中旬，国内疫情形势最为严重，且这个期间的病毒传播速度极快，迅速蔓延至全国。进入 3

月份后，疫情传播得到有效控制，逐步趋于稳定。对于个省市的情况，除湖北之外，广东、河南和浙江的疫情最为严重。为了验证武汉 “封城” 措施的必要性和合理性，通过对武汉人口迁徙强度 (包括武汉人口迁出目的地的人口比例和迁入武汉的来源地的人口比例) 分析得出，平均人口迁徙强度越大，14 天后 (迁徙发生 14 天后) 的累积确诊病例数明显上升。同时，本文还从数理的角度，进一步地证实了平均人口迁徙强度与 14 天后的累积确诊病例高度相关 (Pearson 相关系数为 0.89 (省外)，0.93 (省内)；距离相关系数为 0.89 (省外)，0.90 (省内))。因此，从人口迁徙的角度分析，武汉 “封城” 都是及时且合理的疫情防控措施。

表 8 多源数据融合的约简类 Poisson 回归模型的参数估计及其参数检验

Coefficients	Estimate	Std.Error	z_value	$\Pr(>\|z\|)$
Intercept	−2.505e+00	7.590 e−01	−3.300	0.000976 ***
$cum\_confirm$	3.120 e−03	1.521 e−04	20.506	<2 e−16 ***
cum_dead	−1.661 e−01	1.114 e−02	−14.910	<2 e−16 ***
H	−1.264 e−02	1.888 e−03	−6.696	2.44 e−11 ***
$windv$	−2.862 e−01	5.311 e−02	−5.388	7.53 e−08 ***
$windc$	3.748 e−01	8.329 e−02	4.500	7.00 e−06 ***
P	5.642 e−03	7.393 e−04	7.632	2.89 e−14 ***
$rain$	−2.261 e−02	5.345 e−03	−4.231	2.38 e−05 ***
wu_out	7.916 e−02	2.577 e−02	3.071	0.002148 **
in_wu	9.327 e−02	2.629 e−02	3.548	0.000392 ***
wutod	−1.650 e−03	8.721 e−05	−18.924	<2 e−16 ***

表 9 多源数据融合的约简类 Poisson 模型与全模型的 Wald 检验

模型		残差	自由度	Chisq	$\Pr(>Chisq)$
Model 0: newconfirm	All Variables	3848			
Model 1: newconfirm	$cum\_confirm + \cdots + wutod$	3844	4	3.0845	0.4562

除此之外，本文还发现，单独对气象数据建模，并不能得到合理的、符合客观实际的 Possion 回归模型。但是，在多源数据融合的 Possion 回归模型影响分析中发现，COVID–19 病毒的传播模式与大气压、湿度和风速等气象特征显著相关，但与温度变化显著不相关。通过对这个模型分析还得出，COVID–19 病毒的传播模式也与武汉迁出目的地的人口比例、迁入武汉来源地的人口比例以及武汉与其他城市的空间距离显著相关。接下来，本文的后续工作将逐步增加样本容量，考虑模型的稳定性，进一步研究全球疫情的传播模式以及影响因素，望能对疫情防控提出有价值的建议。

[参考文献]

[1] 中国疾病预防控制中心. 新型冠状病毒肺炎疫情分布 [DB/OL]. http://2019ncov.chinacdc.cn/2019-nCoV/global.html.

[2] Riou J, Althaus C. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020 [EJ]. Eurosurveillance, 2020, 25(4): 2000058. Doi: 10.2807/1560-7917.ES.2020.25.4.2000058.

[3] Giordano G, Blanchini F, Bruno F, et al. Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy [J]. Nature Medicine, 2020, 26: 855–860.

[4] Hao X J, Cheng S S, Wu D G, et al. Reconstruction of the full transmission dynamics of COVID-19 in Wuhan [J]. Nature, 2020, 584(7): 420–424.

[5] Prem K, Liu Y, Russell T, et al. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: A modelling study [RJ]. The Lancet, 2020, 5(5): E261–E270.

[6] Anzai A, Kobayashi T, Linton N M, et al. Assessing the impact of reduced travel on exportation dynamics of novel coronavirus infection (COVID-19) [EJ]. Journal of Clinical Medicine, 2020, 9(2): 601. Doi: 10.3390/jcm9020601.

[7] Wilson N, Kvalsvig A, Barnard L, et al. Case-ratality risk estimates for COVID-19 calculated by using a lag time for fatality [J]. Emerging Infectious Diseases, 2020, 26(6): 1339–1441.

[8] 张久军, 李琦, 杨瑞梅, 何川. 检测对数正态分布位置参数和尺度参数的控制图 [J]. 数理统计与管理, 2018, 37(5): 864–870.

[9] Linton N, Kobayashi T, Yang Y, et al. Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data [J]. Journal of Clinical Medicine, 2020, 9(2): 538–550.

[10] Sun K, Chen J, Viboud C. Early epidemiological analysis of the coronavirus disease 2019 outbreak based on crowdsourced data: A population-level observational study [EJ]. The Lancet, 2020, 2(4): E201–E208. Doi: 10.1016/S2589-7500(20)30026-1.

[11] 李玉莹, 张景肖. 成分数据的 logistic 回归模型研究 [J]. 数理统计与管理, 2019, 38(3): 442–449.

[12] Shen C. Logistic growth modelling of COVID-19 proliferation in China and its international implications [J]. International Journal of Infectious Diseases, 2020, 96(7): 582–589.

[13] Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China [J]. The Lancet, 2020, 395(10223): 497–506.

[14] Bragazzi N, Dai H J, Damiani G, et al. How big data and artificial intelligence can help better manage the COVID-19 pandemic [EJ]. International Journal of Environmental Research Public Health, 2020, 17(9), 3176. Doi: 10.3390/ijerph17093176.

[15] Iqbal M M, Abid I, Hussain S, et al. The effects of regional climatic condition on the spread of COVID-19 at global scale [EJ]. Science of the Total Environment, 2020, 739(10), 140101. Doi: 10.1016/j.scitotenv.2020.140101.

[16] Szekely G, Rizzo M, Bakirov N, et al. Measuring and testing dependence by correlation of distances [J]. The annals of statistics, 2007, 35(3): 2769–2794.

[17] 贺建风, 付永超, 熊健. 基于分层贝叶斯广义线性模型的小域估计方法研究 [J]. 数理统计与管理, 2019, 38(2): 247–260.