Journal of Data Science 18 (5), 849–859 DOI: 10.6339/JDS.202012_18(5).0002
December 2020

On Identification of High Risk Carriers of COVID-19 Using Masked Mobile Device Data

Da Huang[1], Xuening Zhu[2], Weidong Luo[3], Hao Yin[3], Jing Hong[3], Yu Chen[4], Jing Zhou[5], and Hansheng Wang*[4]

[1]*School of Management, Fudan University, Shanghai, China*
[2]*Institute for Big Data, Fudan University, Shanghai, China*
[3]*Aurora Moble Ltd., Shenzhen, Guangdong, China*
[5]*School of Statistics, Renmin University of China, Beijing China*
[4]*Guanghua School of Management, Peking University, Beijing China*

Abstract

Millions of people travel from Wuhan to other cities from Jan. 1st 2020 to Jan 23rd 2020. Taking advantage of the masked software development kit data from Aurora Mobile Ltd and open epidemic data released by health authorities, we analyze the relationship between number of confirmed COVID-19 cases in a region and the people who traveled from Wuhan to this region in this period. Further, we identify high risk carriers of COVID-19 to improve the control of COVID-19. The key findings are three-folds: (1) in each region the number of high-risk carriers is highly positively correlated with the severity of illness; (2) history of visit to the 62 designated hospitals is the foremost index of risk; (3) the second most important index is the travelers' duration of stay in Wuhan. Based on our analysis, we estimate that, as of February 4, 2020, (a) among the 8.5 million people held up in Wuhan, there are 425 thousand high risk carriers; and (b) among all the 3.5 million migrant workers held up in Hubei, there are 175 thousand high risk carriers. The disease control authorities should closely minotor these groups.

Keywords *mobile device software development kit; risk scoring; SARS-CoV-2*

*Corresponding author. Email: hansheng@gsm.pku.edu.cn.

Received March, 2020; Accepted April, 2020

Journal of Data Science 18 (5), 849–859
2020 年 12 月
DOI: 10.6339/JDS.202012_18(5).0002

关于利用移动设备脱敏数据识别新型冠状病毒传染高风险人群的探索性研究

黄达[1]，朱雪宁[2]，罗伟东[3]，殷浩[3]，洪晶[3]，陈昱[4]，周静[5]，王汉生*[4]

[1] 复旦大学管理学院
[2] 复旦大学大数据学院
[3] 极光大数据公司
[5] 中国人民大学统计学院
[4] 北京大学光华管理学院

摘要

本文利用极光大数据公司特有的开发工具包脱敏数据以及网络公开疫情数据，分析了 400 余万从 2020 年 1 月 1 日至 2020 年 1 月 23 日期间，从武汉流出到各地的人员。尝试建立他们与流入区域新型冠状病毒肺炎确诊人数之间的相关关系，并以此识别高风险人群，为各地疫情管控工作提供参考。我们的分析发现：（1）流入各个地区的风险携带者总量同该地区的疫情严重程度高度正相关；（2）是否在高风险期曾经在武汉市 62 家设有发热门诊的指定医院到访过是判断个体风险强度最重要的指标；（3）高风险期在武汉的累计滞留时长是判断个体风险强度第二重要的指标。由此，我们估算，截止于 2020 年 2 月 4 日：（1）被困武汉城内的 850 万人口中，还有大约 42.5 万高风险携带者；（2）被困湖北的 350 万外地务工者，大约有 17.5 万高风险携带者。这些人群是各地疫情管控的重点对象。

关键词 风险得分；移动设备开发工具包

1 引言

从 2019 年 12 月底开始，新型冠状病毒（SARS-CoV-2）肺炎，简称新冠肺炎，在湖北省武汉市爆发，疫情发展迅速，数天内就扩散到全国乃至世界各地。为了控制疫情的进一步恶化，有关方面决定于 2020 年 1 月 23 日上午 10 时起武汉实施封城，各地方政府也相继启动重大突发公共卫生事件一级响应。世界各国对疫情高度重视，纷纷出台相应的旅行限制与公共卫生防护措施。2020 年 1 月 30 日, 世界卫生组织（WHO）宣布，将新冠肺炎疫情列为“国际关注的突发公共卫生事件”。

在疫情防控工作中，对发病人数等关键指标进行建模与预测，无疑是一个重要的任务。对这些指标分析得到的结论，将直接决定应该采取什么样的防控措施，以及需要多大的投入力度。相关问题的研究是目前疫情防控工作的重点，也到了学界的广泛重视。研究传染病问题的常用模型有 SIR(马知恩等, 2004)、SEIR(马知恩等, 2004; Chen et al., 2020)，SEIJR(刘畅等, 2004)，等等。对于这些模型的系统性介绍可以参见近期专著 (马知恩等, 2004; Li, 2017; Ma et al., 2009; Vynnycky et al., 2010)。最近严阅等 (2020) 则提出一类基于时滞动力学系统的传染病动力学模

*通讯作者。电子邮箱: hansheng@gsm.pku.edu.cn。

投稿于 2020 年 3 月; 接收于 2020 年 4 月

型来刻画卫健委公布的累计确诊数和治愈数的动态关系。这些研究所使用的数据均源自卫健委的发布，而后者存在着重大缺陷 (Zhao et al., 2020)。从上述已经发表或公开的文献中不难看出：首先，分析对象单一化，这些模型往往只限于不同时间点上各种类型的人数，而忽视了其他有用的信息；其次，由于疫情发展迅速，以武汉为代表的核心疫区的医疗资源和社会资源均已达到极限，数据无法及时汇总，在时效上存在着严重的滞后；再次，在初期以核酸试剂检验结果是否为阳性作为确诊的唯一依据，而合格的试剂与检验人员数量有限，实际上卫健委所发布的数字严重低于真实的情况。加之各个厂家研发的试剂质量良莠不齐，检验过程复杂，存在的变数多，出现了大量的假阳性或假阴性的诊断结果，即便公布的数据也不尽然准确。考虑到武汉地区的各项关键指标人数均占全国总量的绝对多数，因此依据卫健委发布数据所拟合的模型，不论是估算出来的参数，还是后续的预测，均可能与实际情况存在很大的偏差。综合上述考虑，对于疫情的建模分析不应拘泥于人数问题，也不应局限在既有的模型。

就非武汉地区来说，对区内发病人数等指标的分析和预测，仅仅是从总量的角度的研究。在各地医疗与社会资源都极度紧张的情况下，如能利用其他相关数据，筛选出重点关注人群，则无疑将会显著提高工作效率。事实上，从 2020 年 1 月 1 日起，截至 2020 年 1 月 23 日（武汉封城前），数以百万的曾在此期间滞留过武汉的人群（简称：风险携带者），在并不完全明了疫情严重性的情况下，离开武汉前往全国乃至世界各地，并对流入区域的疫情产生了可见的影响 (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, 2020; Ai et al., 2020; Wu et al., 2020)。毋庸置疑，由于可能和被确诊的患者曾经密切接触等原因，他们是潜在的携带新冠病毒的高风险人群。因此，对非武汉地区来说，辖区内的这批人格外值得密切关注。这里产生了几个重要的问题：(1) 哪些区域有风险携带者流入？(2) 流入的风险携带者都有什么特征？(3) 对流入目的地产生了怎样的疫情后果？这些问题的回答对于接下来各地有关部门的疫情防控工作具有重大参考价值。为了回答这些问题，本文采试图用极光公司相关脱敏数据，并结合网络公开数据源，建立统计学模型。

本文剩余部分结构如下：在第 2 节中具体介绍数据；在第 3 节中通过描述统计，建立对数据的直观认识；第 4 节构建并解释模型；第 5 节利用模型对高风险人群进行识别；第 6 节总结全文，讨论本文的研究意义和局限性，并对未来的研究方向进行展望。

2 数据整理

我们的数据源自卫健委发布的信息和极光大数据公司。极光大数据是目前中国领先的移动大数据服务平台，专注于为移动应用开发者提供稳定高效的消息推送、即时通讯、统计分析、社会化组件和短信等开发者服务。截止到 2019 年 6 月，极光已经为超过 40 万移动开发者和 128.9 万款移动应用提供服务，其开发工具包（SDK）安装量累计 266 亿，月度独立活跃设备 11.3 亿部，具有极好的全网覆盖性。

本文所使用的数据经由移动设备的轨迹数据编码并脱敏后获得，采集时间段为 2020 年 1 月 1 日至 2020 年 1 月 30 日，将该时间段分为前后两个时间窗口：第一个时间窗口为武汉封城前，即 2020 年 1 月 1 日至 2020 年 1 月 23 日，此时间窗口为武汉地区高风险时段；第二个时间窗口为武汉封城后，即 2020 年 1 月 23 日以后，该时间段为从武汉流出到其他地区的时段。我们关注风险携带者，即在武汉封城前曾在武汉滞留过，但在武汉封城后又曾在其他地区出现过的人。

我们认为，该人群具有传播疫情的风险性，他们的流动会对所在地的防疫工作带来极大的影响。根据筛查，风险携带者共计约 400 余万人。需要特别说明的是：极光大数据覆盖的是移动设备并不是自然人。而本文假设每台移动设备对应着一个自然人。这显然是一个为便于分析解读而做出的假设，因为有可能多台移动设备对应一个自然人，也可能一台移动设备对应多个自然人。但是，我们不具备可靠的技术手段予以识别区分。因此，本研究近似地将每台移动设备看作一个独立自然人。这会带来一定的分析偏差，但也许不会影响主要的核心结论。

接下来，我们将全国划分成 444 个省（或直辖市）直管行政区域，简称“直管区域”或“区域”。这包括：（1）86 个直辖市的辖区，（2）333 个地级行政区，（3）25 个省直辖县级行政区。扣除武汉，剩余 443 个直管区域。根据极光数据，这 443 个省直管区域中有 365 个有风险携带者流入（来自 400 余万风险携带者），其中有 354 个爆发了新型冠状病毒感染的肺炎疫情（截至 2020 年 1 月 30 日 24 时）。在爆发疫情的 354 个省直管区域中，疫情的严重程度又各不相同。这就进一步提出了两个相关问题：第一，同样爆发疫情的直管区域，为什么有的疫情特别严重，有的相对较轻？第二，疫情的严重程度与流入的风险携带者数目，以及他们曾经在武汉的活动规律有何关系？

将每个已经爆发疫情的直管区域看作一个样本，并整理出相应的数据，样本量为 354。对每个样本（直管区域），从丁香园网站（www.dxy.cn）获取其累计确诊人数（截至 2020 年 1 月 30 日 24 时），这是因变量。接下来对每个直管区域，根据其流入的风险携带者的各种特征，整理解释性变量。因此，这里的解释性变量都是用于描述直管区域的变量（而不是描述风险携带者的）。但是，它们的计算却是基于风险携带者的自身特征（例如：各种标签），经过各种汇总计算而得。我们在分析的过程中尝试了许多其他的标签或变量，但是综合对比各种方案，最终仅保留了三个变量，其他变量在此不做赘述。具体而言，主要考虑以下解释性变量。

第一、风险携带者总量。这是最重要的一个解释性变量，直接刻划了该直管区域承受的疫情风险强度。这个指标的具体计算细节如下：对每个直管区域计算，在指定时间窗口内，从武汉流入的风险携带者总数（来自那 400 多万风险携带者），并经过必要的标准化和对数变化后获得。因此，该指标的单位不是人数。

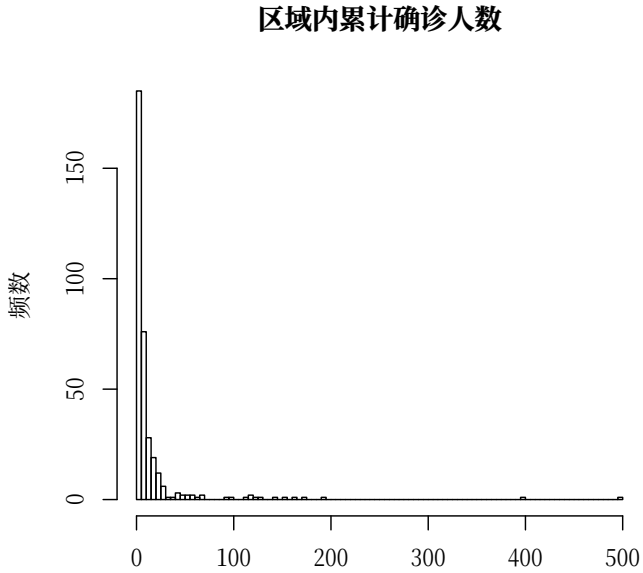
第二、武汉滞留时长。对于一个特定的直管区域，在指定的时间窗口内，所有流入的风险携带者，在高风险时期内，所有人（不是一个人）的在武汉滞留时长总和，并经过必要的标准化和对数变化后获得。因此，该指标的单位不是自然时间。该变量可以从另一个侧面衡量直管区域所承受的疫情风险强度。

第三、定点医院到访比例。对于一个特定的直管区域，在指定的时间窗口内，有多大比例的风险携带者曾经在高风险时期内在武汉 62 家有发热门诊的指定医院中到访过。到访比例越高，则流入人群携带病毒的风险就越高，也越有可能将病毒传给其他人。如果该指标取值为 0 表示直管区域所有流入的风险携带者中没有任何人到访过定点医院。相反，如果该指标取值为 1 则表示直管区域所有流入的风险携带者全部都到访过定点医院。

3 描述统计

我们首先对所有变量做描述统计，以期获得对数据的直观认识。

第一，累计确诊人数。各直管区域内截至 2020 年 1 月 30 日 24 时的累计确诊人数（下称确



区域内累积确诊人数描述性统计

最小值	1
25% 分位数	2
中位数	5
均值	15.3
75% 分位数	11
最大值	496

图 1: 左：区域内累计确诊人数直方图；右：区域内累计确诊人数描述性统计。

诊人数）分布如图 1左侧所示。数据呈现出明显的右偏状态，绝大多数区域只有少量确诊者，少数区域出现集中爆发情况。由图 1右侧的描述统计量可知，所有区域中累计确诊人数最小值为 1（包括重庆市巴南区等 52 个区域），最大值为 496（湖北省黄冈市），平均人数略高于 15 人。

第二，风险携带者总量。所有直管区域内风险携带者总量的分布如图 2左侧所示，其平均值约为 −2.2 个单位，最小为甘肃省临夏市（−5.88 个单位），最大为湖北省孝感市（1.81 个单位）。为了解风险携带者总量与累计确诊人数之关系，以各区域累计确诊人数的 25% 分位数（2）和 75% 分位数（11）为临界点，将所有的直管区域分为高、中、低三组，以柱状图的形式展示三组中平均风险携带者总量，参见图 2右侧。显然，累计确诊人数高的组内，平均风险携带者总量也最高；相对比例为 −0.91 个单位，中组次之，低组最少，相对比例为 −3.18 个单位。这说明该指标与直管区域疫情边际正相关。

第三，武汉滞留时长。图 3左侧显示了各直管区域内所有风险携带者在武汉累计停留的时长分布。请注意本指标是被标准化和对数变换过的，因此没有单位。从直方图可以看出，平均滞留时长大概为 20 个单位。以各区域累计确诊人数的 25% 分位数（2）和 75% 分位数（11）为临界点，将所有的直管区域分为高、中、低三组，以柱状图的形式展示三组中风险携带者在武汉累计滞留时长的平均值，参见图 3右侧。显然，累计确诊人数高的组内，累计滞留时长平均值也最高（平均为 22 个单位）；中组次之，低组最少（平均为 19 个单位）。这说明该指标同直管区域疫情边际正相关。

第四，定点医院到访比例。图 4左侧展示了各直管区域内风险携带者中到访过武汉 62 家指定发热门诊医院的比例分布。图 4左侧直方图显示，平均到访比例大概为 0.033。我们以各区域累计确诊人数的 25% 分位数（2）和 75% 分位数（11）为临界点，将所有的直管区域分为高、中、低三组，以柱状图的形式展示三组中定点医院到访比例的平均值，参见图 4右侧。显然，累计确诊人数高的组内，定点医院的算术平均到访比例均值也最高，平均为 0.042；中组和低组最少，为 0.03。这说明该指标同直管区域疫情边际正相关。

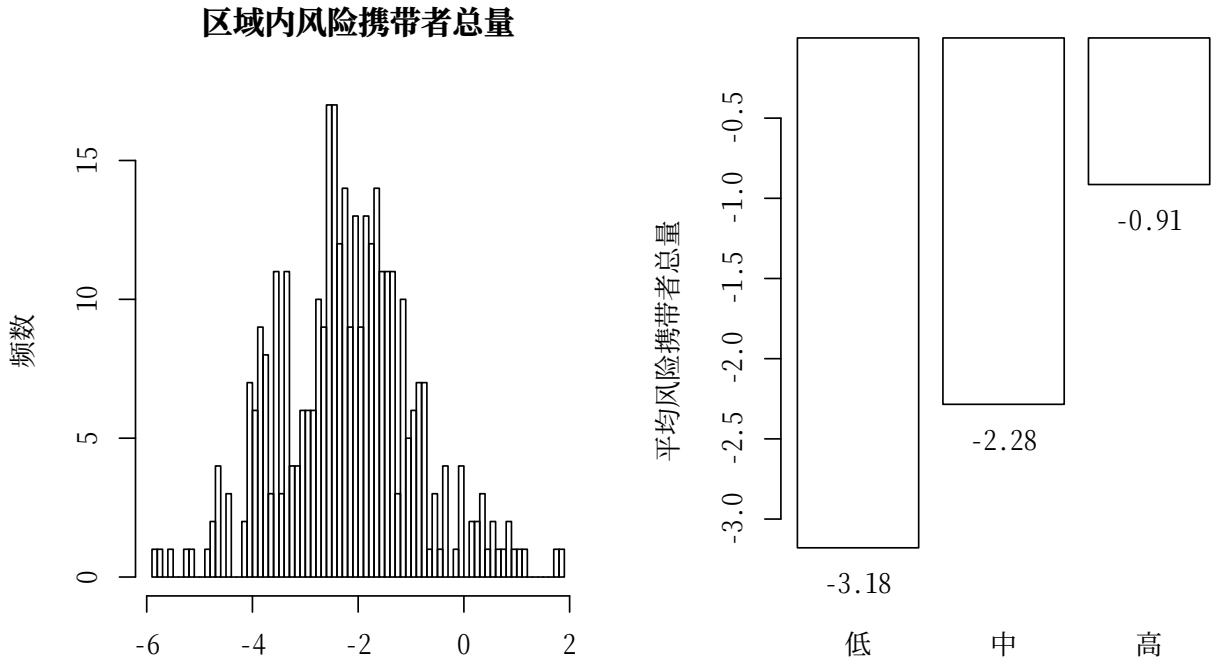


图 2: 左：区域内风险携带者数直方图；右：不同分组区域内平均风险携带者数。

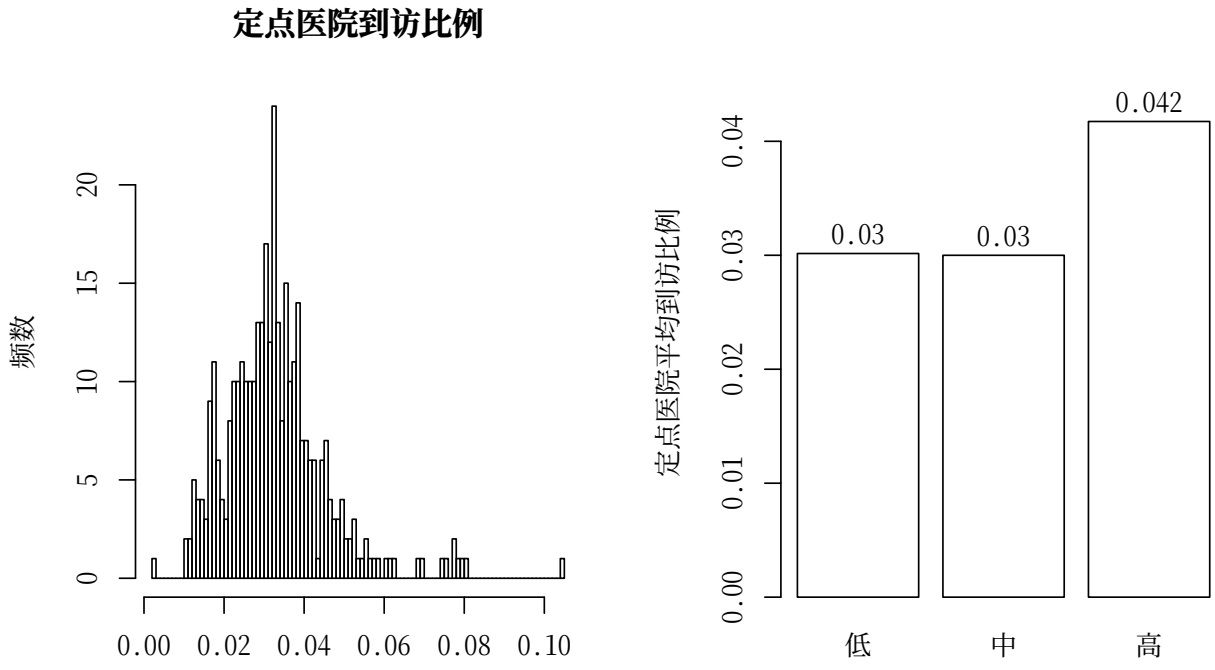


图 3: 左：累计武汉滞留时长直方图；右：不同区域分组累计武汉平均滞留时长。

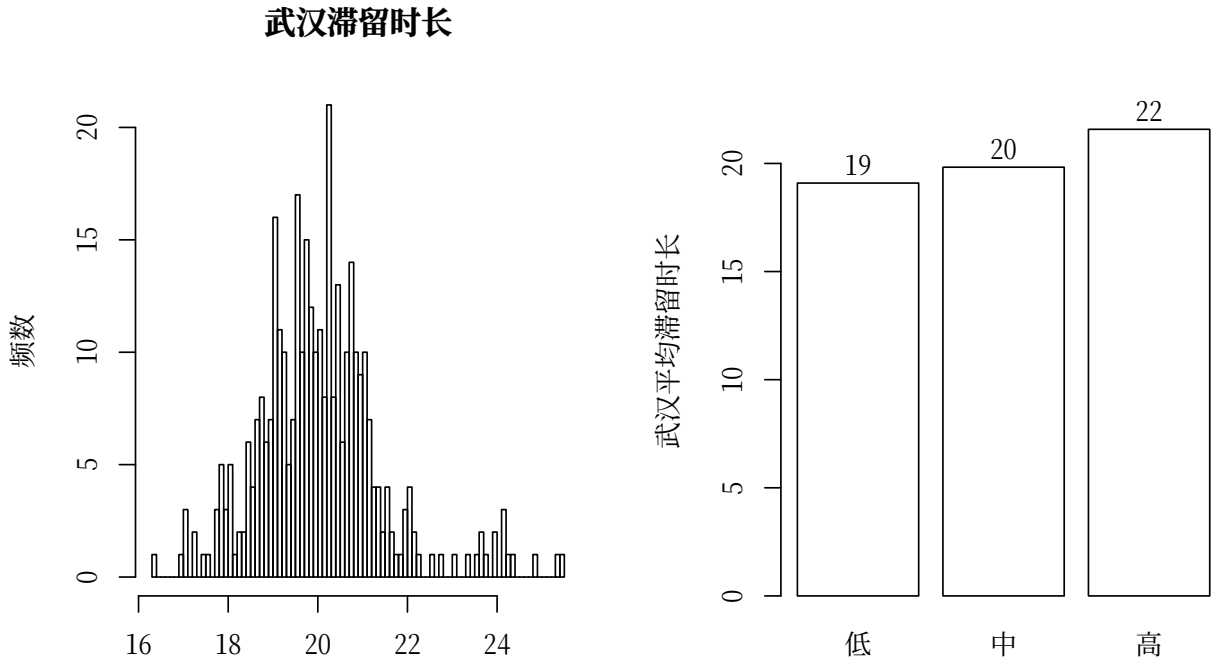


图 4: 左：定点医院到访比例的直方图；右：不同分组的定点医院平均到访比例。

表 1: 回归模型的参数估计结果。

	估计	标准误	t-值	p-值	显著性
截距项	−3.381	1.575	−2.146	0.0325	*
风险携带者总量	0.387	0.075	5.161	<0.001	***
武汉滞留时长	0.261	0.073	3.556	0.0004	***
定点医院到访比例	22.586	3.578	6.312	<0.001	***

4 模型分析

如前所述，我们研究考虑大量的用于描述风险携带者在武汉以及在最终直管区域行为特征的指标。但是，在线性回归的框架下，绝大多数无法通过严格的变量筛选（BIC 准则）。以对数累计确诊人数为因变量，最终能够通过模型检验的变量和模型拟合结果如下：

模型的 R^2 为 0.6323，调整后的 R^2 为 0.6292，这表明模型的拟合优度良好。对模型做了 F-检验与 t-检验，结果表明模型与所有参数都高度显著。我们还对模型作了基本的残差分析，未发现重大模型设定问题。从表 1中可以看到：（1）直管区域风险携带者总量越多，则确诊人数越多；（2）直管区域风险携带者中，累计武汉滞留时长越长，则确诊人数越多；（3）直管区域风险携带者中，到访过武汉 62 家定点医院的到访比例越高，则确诊人数越多。

5 高风险人群识别

根据上述分析结果可知，影响风险大小的个体因素主要有两个：高风险时期在武汉的滞留时长（回归系数：0.26）；高风险时期是否曾在武汉 62 家有发热门诊的指定医院中到访过（回归系数：22.59）。这提示我们可以构造一个简单的面向个体的风险打分方法，公式如下：

$$\text{风险得分} = 0.26 \times \text{武汉滞留时长} + 22.59 \times \text{是否到访过 62 家定点医院} \tag{1}$$

其中，第一个变量（武汉滞留时长）代表该个体在武汉的滞留时长（经过同样标准化和对数变换）；如果曾到访过 62 家定点医院，则第二个变量（是否到访过 62 家定点医院）取值为 1，否则取值为 0。该风险得分的分值可以被标准化成 0 到 1 之间的一个数，但为了避免与我们常用的概率混淆，我们保留上述未标准化的版本。

值得注意的是，本公式中的权重来自于回归分析（表 1）。而该回归分析的基本单位是直管区域，而非自然人个体。因此，该打分方法一定不是最优的。但是，这是目前我们能够想到的最简单打分方法，该得分有助于快速识别相对的高风险携带者。尽管个体得分与其发病概率之间的数量关系并不清楚，但我们认为快速（但不是绝对精确）识别高风险携带者仍具有很大的现实意义。主要原因有两个：

第一、为当前各地防疫工作建议重点方向：在当前社会与医疗资源高度紧张的情况下，对已经流入各地的 400 余万风险携带者以及目前滞留武汉城内的 850 万人群，进行快速排查，识别其中的高风险携带者，集中力量进行重点防控，能够极大提高疫情防控效率，有效避免疫情的进一步扩散。

第二、助力疫后经济与生产的恢复：基于初步估算，大约有 350 万湖北以外务工的人员被困湖北。随着返城浪潮的开始，各地逐步解禁，他们会再次流动到全国各地复工。对于流入地来说，尽可能精准识别其中的高风险携带者，不仅能够有效提高防疫工作效率，也可以让大多数低风险者迅速复工，恢复生产。

基于该公式，我们对已流出武汉的 400 余万人给出风险得分，并根据其在武汉期间有没有到访过发热定点医院分成两类，其中没有到访过的人占 94.24%，到访过的人占 5.76%。对两类人按照得分大小分别做饼图，参见图 5。可见到访过医院的人（右侧，得分都在 20 以上）分值要远远高于没有到访过医院的人（左侧，得分一般不超过 5）。由此可见是否到访过定点医院是识别风险携带者的最主要指标。

值得一提的是，在整个分析中，我们还考查了风险携带者滞留武汉期间的各种行为特征。例如：是否到访过华南海鲜市场，等等。但是，这些指标都经不起严格的模型筛选。在我们分析过的所有变量中，最能经得起挑战的就是定点医院到访比例。这说明，是否到访过定点医院是识别高风险人群最重要的一个特征。综合以上分析，具体建议如下：

第一，在当前防疫工作中，武汉需要特别确认被筛查对象在高风险期内有没有到访过这 62 家发热定点医院。如果到访过，需要特别注意，并结合医学检验手段予以确证。目前粗糙估计滞留武汉城内的人群有 850 万人以上，其中到访过定点医院的人群占比大约 5%。这意味着武汉城内还有大概 42.5 万高风险携带者需要密切关注。

第二，在节后返程大潮中，对于湖北省流出的人群，同样需要专门确认是否曾经到访过这些

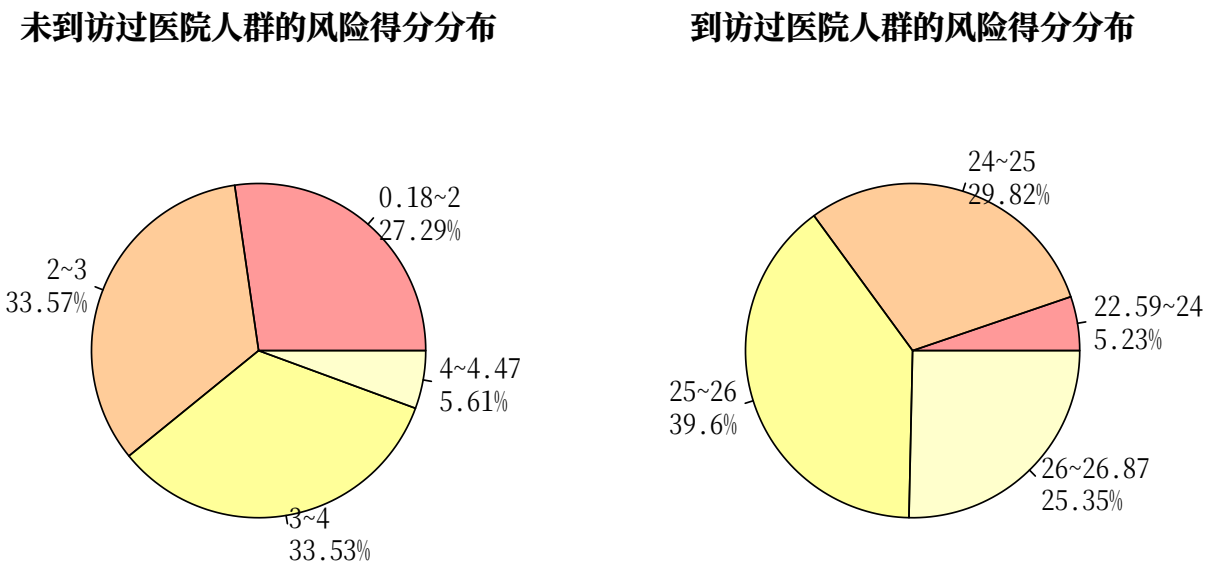


图 5: 左：未到访过医院的人得分饼图；右：到访过医院的人得分饼图。

定点医院。目前初步估计滞留湖北的外地（湖北以外）务工人员大约 350 万。如果也按照大概 5% 的比例推算，大概对应着 17.5 万人要在逐步离开湖北前往全国各地，这也是值得密切关注的人群。

6 总结与讨论

本文利用极光大数据公司独特的 SDK 脱敏大数据对年内曾在武汉滞留，但现已流散在全国各地的 400 余万风险携带者做了统计分析，研究了他们的各种特征与所在直管区域疫情的相关关系，确立了三个强相关指标：（1）风险携带者总量；（2）累计武汉滞留时长；以及（3）定点医院到访比例。通过回归分析发现，定点医院到访比例是识别高风险携带者的最有效的特征。基于该发现，我们对武汉城内的高风险携带者总数做了初步估计，也对滞留湖北即将返回全国各地的高风险携带者总数足了作了估计。

本文的分析结果有很多局限性，因此所呈现的结果需要批判性解读，审慎采纳。本研究主要的缺陷如下：

（1）数据采集区间。受制于时间限制和数据匹配等各种现实原因，本研究所采用的数据并不是最新数据，而是截止到 2020 年 1 月 30 日 24 时。而疫情一日一变，防控措施如定点医院设置等也在不断变化，如果用最新数据分析，模型权重必然会改变。但是，也许核心的定性结论（例如：定点医院到访比例的重要性）也许不会改变；

（2）数据质量。本文因变量确诊总数来自丁香园官网。丁香园数据来自于各地卫健委网站和媒体。由于多种原因，包括国家卫健委网站上公布的数据有时都需要推敲，更何况媒体，因此可以预期我们采集的确诊总数也一定存在误差。同时，各自变量的采集也受到设备的开机与否、信号质量、网络延迟等多种因素的影响，无法做到绝对准确。但这可能是目前能够获得的最好的数据，我们也相信这不会影响核心结论。

（3）数据有限。由于时间紧迫，本文无法及时综合所有来源的数据和信息，而且为了防控疫

情，各地社会资源均处于非常紧张的状态，数据与资料无法及时向公众公布。因此，我们的研究仅限于极光大数据公司所提供的信息与已公布的疾控数据。毫无疑问，如能获取更多数据，将会极大推进我们的工作。但我们相信本研究的及时反馈将对接下来的防疫工作，及将来的进一步分析打好基础。

和现有的工作相比，本文所研究的问题不再局限于传染病人数的动态变化，而更关注有哪些因素影响了疫情的变化，为疫情防控中各地政府决策提供所急需的支持。在分析的数据上，也从原来卫健委单一来源，拓展到多源大数据。循着这一方向，有待探索的研究很多，特别是如果可以获得传染病学的调查数据 (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, 2020)，并将其与现有的数据结合起来，必将能够得到更多有用的结论，为进一步的疫情防控和将来的防疫工作奠定坚实的基础。

在本文的基础上，还有许多值得拓展的方向。例如，从地域间的连接上，本文主要考虑的是人口流动对疫情所带来的影响，这种流动主要是由区域间经济往来所形成的。除此之外，我们也注意到疫情在各个区域的爆发程度与地域之间的地理相关性有关。例如，在湖北临近武汉的区域，可以看到显著较高的确诊人数。但同时我们也注意到，离武汉较远的浙江省也出现了疫情爆发的现象。通过查阅相关新闻与资料并研判，我们认为，这是因为浙江省的部分区域与武汉有着高度经济、商业与人员往来导致。这一结论与最终模型中的经过筛选后得到的风险携带者总量和武汉滞留时长这两个变量可以互相印证。因此，如何将地域之间不同的连接信息纳入模型，是未来可进一步进行深入挖掘的课题。

致谢

谨以此文向所有奋战在救治与防疫第一线的英雄们，致以我们最崇高的敬意！

感谢特刊主编和审稿人对我们文章专业的指正与建议，有效地提高了本文的质量与行文规范。

本项研究得到国家自然科学基金项目“基于分布式数据和高维数据的极值统计研究”（基金号：11971115），“高维大数据的若干统计推断问题研究”（基金号：11971116），“电子商务环境中定向广告的精准投放与管理策略研究”（基金号：71531006），“双边市场的动态竞争与生态协调”（基金号：71672042），“大规模网络数据统计建模及高效计算”（基金号：11901105），“时空数据建模和预测研究”（基金号：71991472），“基于超算的大数据分析处理基础算法与编程支撑环境”（基金号：U1811461），“多源异构数据的融合、特征提取与分析方法”（基金号：11831008），“高维复杂数据的理论与应用”（基金号：11525101），“大数据驱动的管理决策模型与算法”（基金号：71532001），国家重点研发计划“空气质量统计诊断模型”（批准号：2016YFC0207704），全国统计科学研究项目“含移动平均成分的多元时间序列降维研究”（立项编号：2015LY77），上海市科技人才计划“大规模网络数据自回归建模及动态分析”（批准号：19YF1402700），泰康溢彩公共卫生及流行病防治专项基金，以及复旦新再灵大数据联合实验室的资助。

参考文献

严阅, 陈瑜, 刘可伋, 罗心悦, 许伯熹, 江渝, 程晋, 2020. 基于一类时滞动力学系统对新型冠状病毒肺炎疫情的建模和预测. 中国科学: 数学, 50(3): 385-392.

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, 2020. 新型冠状病毒肺炎流行病学特征分析. 中华流行病学杂志, 41(2): 145-151.

刘畅, 丁光宏, 龚剑秋, 王凌, 程珂, 张迪, 2004. SARS 爆发预测和预警的数学模型研究. 科学通报, 49: 2245-2251.

马知恩, 周义仓, 2004. 传染病动力学的数学建模与研究. 北京: 科学出版社.

Ai S, Zhu G, Tian F, Li H, Gao Y, Wu Y, Liu Q, Lin H, 2020. Population movement, city closure and spatial transmission of the 2019-nCoV infection in China. medRxiv preprint: https://doi.org/10.1101/2020.02.04.20020339.

Chen T, Rui J, Wang Q, Zhao Z, Cui JA, Yin L, 2020. A mathematical model for simulating the transmission of Wuhan novel Coronavirus. bioRxiv preprint: https://doi.org/10.1101/2020.01.19.911669.

Li MY, 2017. An introduction to mathematical modeling of infectious diseases. Springer.

Ma S, Xia Y, 2009. Mathematical understanding of infectious disease dynamics. World Scientific Publishing.

Vynnycky E, White R, 2010. An introduction to infectious disease modelling. Oxford University Press.

Wu JT, Leung K, Leung GM, 2020. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. The Lancet, 395(10225): 689-697.

Zhao S, Musa SS, Lin Q, Ran J, Yang G, Wang W, Lou Y, Yang L, Gao D, He D, Wang MH, 2020. Estimating the unreported number of novel coronavirus (2019-nCoV) cases in China in the first half of January 2020: A data-driven modelling analysis of the early outbreak. Journal of Clinical Medcine, 9(2): 388.