·疫情预警与防控·

基于人口流动的新冠肺炎疫情风险分析

贾建民[1*] 袁 韵[2] 贾 钰[3]

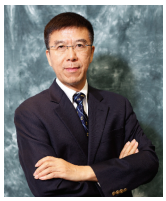
1. 深圳高等金融研究院/深圳市人工智能与机器人研究院/香港中文大学(深圳) 经管学院，深圳 518172
2. 清华大学 经济管理学院，北京 100084
3. 香港大学 经济与工商管理学院，香港特别行政区 999077

[摘 要] 本文报告我们承担的国家自然科学基金新冠专项项目的研究进展，其主要结果已发表在 *Nature*(Jia 等人[1])期刊，同时也报告一些相关的扩展分析。首先，基于在疫情爆发前武汉输入到全国各地的人口数量，建立了时空基准风险模型，成功地预测和解释了新冠疫情在全国各地的时空分布特征。进一步，为了在新冠疫情初期对不同地区的疫情风险进行评估，我们发展了一个社区传播风险指数以及风险探测的一套工具，而传播风险是根据实际确诊人数偏离基准模型预测的显著程度来测量的。这一指数可用于疫情预警系统，来辨识和追踪那些具有高传播风险的地区。最后，我们用统计模型和机器学习的方法评估各种人口流动风险源的异质性，检验他们对新冠疫情传播的相对贡献大小，包括比较武汉居民与非武汉居民、不同年龄段、不同性别等对疫情传播的影响。

[关键词] 人口流动；新冠肺炎；疫情传播；时空风险模型；风险分析

此次新冠肺炎(COVID-19)疫情是新中国成立以来在我国发生的传播速度最快、感染范围最广、防控难度最大的一次重大突发公共卫生事件。作为一种新发重大传染病，新冠肺炎的病理和传播特征在其疫情初期不为研究人员所熟知，从而影响了对其的监测、预警和应对。因此，此次疫情对我国医疗卫生体系提出了重大挑战，也对我国经济社会造成较大冲击。若能总结新冠肺炎疫情防控阻击战中的经验，围绕新发重大传染病的疫情防控应对与管理，开展基础性、回顾性、实证性和前瞻性的研究，则可为打赢疫情防控阻击战提供强大支撑，为科学防控和应对疫情等重大突发公共卫生事件、减轻其对我国经济社会的影响、完善国家治理体系和提升社会管理能力提供决策支撑和对策建议。

传染病疫情的发展是通过人们的流动与接触来传播的。利用个体层面的手机位置数据可以很好地监测大规模人类迁徙，其优势包括两方面：(1) 手机位置提供了关于手机用户移动的实时、精确和可验证的数据；(2) 我国已有超过 10.8 亿手机用户，手机覆盖人口占总人口比例极高。合理利用个体层面的手机位置数据，把握地区间人口流动的数量和方向，可以迅速和准确地追踪潜在感染人群的流动情况，有助于政府防疫部门实时精准监测和预警，特别是在疫情爆发初期就能准确预测疫情发展的地点、



贾建民 香港中文大学(深圳)校长讲席教授，曾任香港中文大学商学院市场学系教授兼系主任，西南交通大学经济管理学院院长，教育部长江学者讲座教授，全国 MBA 教育指导委员会委员，国家自然科学基金委管理科学部专家咨询委员会委员，美国营销科学学会(MSI)学术董事，美国 *Operations Research* 杂志副主编。国家杰出青年科学基金获得者，教育部创新团队负责人，2019 年“复旦管理学杰出贡献奖”获得者。主要从事大数据分析、市场营销、决策分析、风险分析等方面的研究。

收稿日期：2020-11-18；修回日期：2020-12-09

* 通信作者，Email：jmjia@cuhk.edu.cn

本文受到国家自然科学基金项目(72042009 和 72074072)、深圳市人工智能与机器人研究院(2020—NT001)的资助。

规模以及时间长短，为政府和医疗机构在应急响应、资源分配和公共卫生决策等方面提供重要依据。对我国这样一个人口众多但医疗资源有限的发展中国家而言，基于时空大数据实现迅速和准确的新发重大传染病监测、预警和应对尤为重要。

对于这次新冠肺炎疫情爆发，充分显示了伴随着社会经济发展而产生的大规模人类迁徙极易将疫情扩散到广泛的地区，使得突发公共卫生事件迅速地从偶发、局部层次升级为大流行。新冠疫情的大流行，给世界各国带来了极其严重的影响，截止 12 月 24 日，全球的感染人数超过 7 000 万，死亡人数超过 170 万，并且疫情还在进一步恶化。由于缺少新冠疫苗等医疗手段，各国只能依靠各种非药物干预手段，例如隔离、封城、限制交通、停工、停学、社交距离等防控措施。我国以及世界各国的抗疫经验充分表明，只要能有效地采取这些防控措施，疫情就能得到抑制。

本文以 Jia 等近期发表在 *Nature* 上的论文[1]以及相关文献为基础，从人口流动性的角度介绍基于时空大数据的新冠肺炎疫情风险评估和监测的相关研究工作。首先，基于在疫情爆发前武汉输入到全国各地的人口流动数据，并与疫情发展的空间地域分布和时间变化趋势关联起来，从而建立了全国各地疫情发展的基准风险模型，其结果可用于预警各地的疫情风险大小，在疫情发展的早期可为应急计划和相关决策提供依据。进一步，我们建立疫情传播风险指数，用以监测各地区疫情的发展是否显著偏离疫情基准趋势，并判断出现社区传播风险的可能性(95%置信区间)，以此作为地区疫情管控的监测指标。最后，采用模型和机器学习的方法对各种从武汉以及湖北流出的人口对全国各地疫情的影响进行风险评估，为今后的应急管理工作提供决策依据。

1 国内外研究现状

智能手机、GPS、App、移动互联网与物联网等现代信息技术的快速发展使收集人们移动轨迹的数据成为可能，从而推动了基于时空大数据开展人群流动性行为的研究[2]。Brockmann 等开创性地利用美国银行票据记录网站对超过 100 万人的票据位移数据开展研究，发现票据位移的距离近似按照幂律分布衰减[3]。Gonzalez 等利用手机通讯数据也证实了手机用户的移动距离服从具有指数尾的幂律分布，而且人们的位移具有简单的再生特征，即经常来回于少数几个近距离地点(例如家和工作地)，只是偶尔有些长距离的旅行[4]。流动性的这些统计特征和周期规律使人们的流动行为具有高度的可预测性[5]。Deville 等进一步研究了人们的流动性与社会网络之间的关系，建立了流动性幂律与互动性幂律两个模型的指数之间的线性关系，揭示了人们的流动性和社会互动的空间依赖性[6]。

现代交通的发展使得在局部出现的流行病随着四通八达的交通网络，尤其是航空系统，在短时间内扩散到世界其他地区，甚至导致大流行病的发生。Eubank 等基于 agent 模型对流行病在区域内的传播进行了研究，生成数百万个体来模拟城市动态、交通及人类活动，获得了较为真实的结果[7]。Colizza 和 Vespignani 构建了人口迁移模型，基于扩散方程推导流行病随交通网络移动传播的动力学机理[8]。

采用人口流动数据和网络分析的方法来预测传染性疾病的传播已经成为流行病领域的重要方法和手段[9-11]。受益于人口流动的轨迹大数据的可获得性和当下越来越强的计算能力，依据社会人口移动分布构建的传播模型能获得更好的预测效果[12]。Bansal 等指出大数据革命将极大提高现有流行病学信息的可粒度和及时性[13]，通过追踪用户手机日志、互联网搜索等踪迹并与传统检测系统结合，可以构建更为准确的传染病预测模型[14, 15]。另外，当公共卫生事件爆发时，引入机器学习算法对数据流进行训练，结合复杂统计方法，可以突破传统预测模型的瓶颈，给出更为有效的预测结果[16]。

除了人口的流动性之外，社会网络也是传播疫情的重要渠道[17, 18]。虽然一些疫情可以通过偶然接触传播，但具有社交关系的人之间的接触时间通常比陌生人之间的接触时间要长得多，因此传染往往发生在家人与朋友之间。Althouse 等利用社交媒体数据并结合网络搜索，将其纳入公共健康评估框架，有效地增强了公共卫生监测的能力[19]。Reich 等通过 Twitter 网络数据和 Google 搜索，构建了具有多年数据比较的流行病学传播地域分布模型，取得了更加满意的预测精度[20]。

在此次新冠肺炎疫情全国爆发的初期，Wu 等采用航空票务订购数据和腾讯在疫情前一年的流动性数据预测了武汉和国内主要城市以及周边一些国家和地区的疫情状况[21]。Li 等同样是基于疫情前两年腾讯的春运流动数据分析了大量没有在案的感染人群推动了这次新冠肺炎疫情的快速传播[22]。但采用过去的航班或者人口流动数据，其研究结论

的准确性值得商榷。Chinazzi 等利用国际航空数据以及百度流动数据，显示了武汉封城以及其他国家的旅行限制很大程度上减轻了疫情的发展[23]。Gilbert 等采用国际航空数据对非洲国家疫情爆发情况进行了预测分析[24]。Rader 等的研究显示，越拥挤的城市疫情传播越严重[25]。限制城市内和城市之间的流动性成为抗击新冠疫情最主要的非药物干预手段[26-28]。

以上分析表明，采用人口流动数据以及社交网络数据等来研究传染病的传播已成为本领域的重要方法和手段，特别是关于新冠肺炎疫情传播的研究。基于这些数据的方法各有利弊。社交媒体和搜索数据只反映了人们对传染疾病的网络态度且受到诸多因素的影响，并且难以开展因果分析。基于人口流动数据的研究更加能够揭示流行病传播的路径和规律，但较难获取能够及时反映疫情传播的流动数据。由于缺乏及时和完整的人口流动数据，以往对疫情爆发期间的流行病学建模通常高度依赖对流动性的统计估计（例如，基于平时航班数量进行预测），而不是针对实际的人口流动进行定量观察和分析。统计估计所获结果常常存在很大偏差，不能有效反映事实上的人口流动数量、速度和人类活动。历史迁移数据是一种可能的替代数据，此次新冠疫情的若干研究使用了历史迁移数据（如前几年的春运迁移）来预测病例增长情况。然而，历史数据在当前的防疫隔离政策下可能已经失去了参考性。不管是统计估计还是历史数据，由此带来的即使是稍微不准确的预测也可能使政府决策产生重大偏差和严重后果：反应不足可能导致疫情蔓延，而反应过度可能导致社会和经济发展蒙受巨大损失。

本项研究直接采用运营商提供的在封城前自武汉流向全国各地市的人口数据，可以准确地描绘疫情的传播和分布规律，因此能够更加有效地监测和预警疫情、提前提出应对措施。

2　疫情发展的时空规律

此次新冠肺炎疫情首先在武汉爆发，并迅速扩散到全国。由于疫情爆发在农历春节前夕，正值大规模的春运人群迁徙，追踪源自武汉及湖北的迁徙人群的流动极为紧迫。而且武汉是我国铁路和航空网络的重要中心枢纽，因此疫情扩散的潜在规模和范围极大。

本项研究将封城之前由武汉流向全国各地级市的人口规模分布与全国各地感染的新冠肺炎病例数量建立起联系，发现其地理空间分布的基本特征可以概括为如下两点：(1)武汉及湖北周边与其有紧密地缘关系的省市感染程度最严重，依次为河南、湖南、安徽、江西、重庆、四川等；(2)与武汉及湖北有紧密社会经济关联的沿海发达省市也面临高度风险，包括广东、浙江、江苏、山东、北京、上海、福建等。武汉疫情爆发以来，这十三个省市的疫情严重程度一直居于全国前列。这种地缘关系和社会经济关联方面的综合强度可以通过武汉及湖北与这些省市之间的人口流动得到反映。

本项研究使用电信运营商提供的全国范围内通过整合个体层面的手机位移数据来追踪武汉封城前流入各地级市的人口数量。相关的人口统计和经济数据来源于 296 个地级市（不包括武汉），其所覆盖行政区域的平均人口为 440 万，占全国总人口的 94.07%。通过对于数据的早期观察，我们发现武汉流入全国各地的人口数量与当地确诊感染人数之间的相关性越来越高，到 2 月中旬以后这一相关系数稳定在 0.95 以上，图 1 显示了这种相关性的变化趋势。而各地的人口数量、GDP 以及百度搜索指数与当地确诊感染人数的相关系数却是随时间推移而逐渐减小。因此，自武汉流入各地的人口分布决定着全国各地最终的疫情分布。

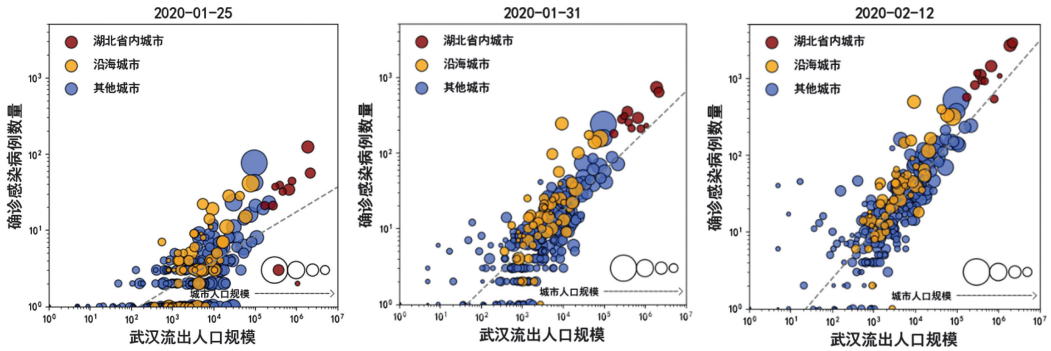


图 1　武汉流向全国各地市的人口数量与当地确诊感染人数之间的相关性变化

Jia 等基于 Cox 比例风险模型(Cox proportional hazards model)框架,建立了多个整合疫情时空发展的风险模型[1]。这里我们介绍一个指数类的模型,其中疫情在时间维度上的发展规律采用 S 形增长的 Logistic 函数来描绘,而疫情在地域上的分布规律按自然指数乘积模型来考虑:

$$\lambda(t \mid x_i) = \frac{\alpha}{1+\mathrm{e}^{-\lambda t+\omega}}\left(\prod_{j=1}^{m}\mathrm{e}^{\beta_j x_{ji}}\right)\mathrm{e}^{\sum_{k=1}^{n}\lambda_k I_{ik}} \qquad (1)$$

式中,$\lambda(t|x_i)$是给定 i 地区自武汉流入人口数量以及其他相关变量条件下在时间 t 时的确诊病例数,$x_i=\{x_{1i},x_{2i},\cdots,x_{mi}\}$,其中 x_{1i}是由武汉流向城市 i 的累积人口数量,x_{2i}和 x_{3i}分别是城市 i 的 GDP 和人口规模,m 是变量数,α、γ、ω、β_i 是待估计参数。模型中还可以加入有研究价值的其他变量 x_{ji}。λ_k 是省份固定效应,n 是城市数量,I_{ik}是城市哑变量,如果 $i\in k$(城市 i 属于省份 k),$I_{ik}=1$,否则 $I_{ik}=0$。为消除各个变量 x_{ji}的单位差异、增加可比性,我们对数据进行标准化处理,$x'_{ji}=(x_{ji}-Mean)/Std$。

我们采用 Levenberg-Marquardt(LM)方法来直接估计模型(1),而该方法兼具最速下降法和牛顿法的优势,被广泛应用于解决多种非线性问题[29]。模型估计除了采用武汉流入 296 个地级市的人口数量以外还包括这些地区的 GDP 和本地人口数据,时间窗口从 1 月 24 日到 2 月 19 日共计 27 天,其总样本量为 $n=7\,992$(Jia 等[1])。首先,只使用自武汉流入各城市人口数量进行拟合,其 $R^2=0.926$。随着数据的增加,疫情的时空风险模型拟合效果越来越好,疫情曲线的图像也越来越清晰。图 2 是基于模型(1)描绘的疫情时空发展规律,随着武汉流入人口数量和时间动态变化的三维曲面,数据点代表相应的确诊病例数。

模型(1)不仅可以用来预测各地的确诊病例数,还可以作为对疫情传播进行监测时的基准,用以找出疫情偏离基准的地区。为评估不同地区的新冠肺炎疫情社区传播风险,我们根据实际确诊病例数与模型预测数之间的差异建立了疫情传播风险的度量:

$$\Delta_i = \sum_{t=1}^{T}[\lambda(t \mid i) - \hat{\lambda}(t \mid x_i)] \qquad (2)$$

其中,$\lambda(t|x_i)$是地区 i 在时间 t 的累计确诊病例数,$\hat{\lambda}(t|x_i)$]是由前述的疫情风险模型估计出的地区 i 在时间 t 的预测病例数,T 是总时间(天)。将 Δ_i标准化,即用 Δ_i减去均值再除以标准差,可以形成最终的疫情传播风险指数 $\bar{\Delta}_i$。图 3 展示了部分城市疫情传播风险指数随时间变化的动态特征。

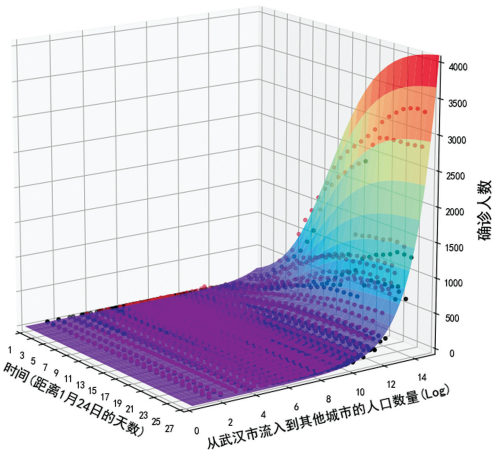


图 2 模型预测与实际疫情(散点)的比较

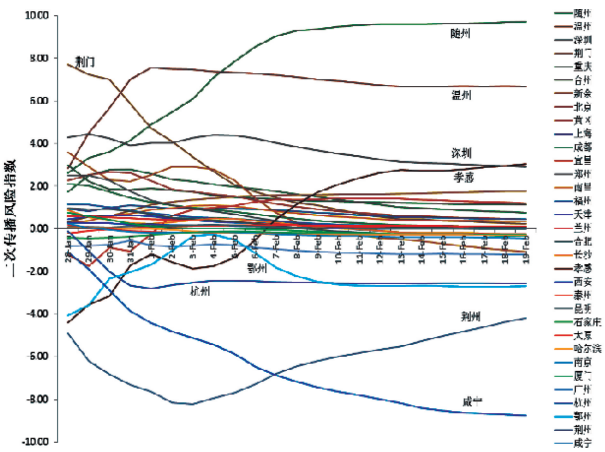


图 3 部分城市疫情传播风险指数的动态变化

在实际应用中,可以把疫情传播风险指数值超过 95%置信上限的地区列为重点监测对象(如 $\bar{\Delta}_i>1.96$)。例如,在 1 月 29 日,该指数表明温州的社区疫情传播风险最严重;当地政府于 2 月 2 日宣布的全面隔离措施证实了模型的预测。到 2 月 19 日,超过 95%置信上限的有随州、温州、孝感和深圳,这些地区产生了相对较高的疫情传播,特别是随州和孝感的传播风险指数随着时间一直在增加。在武汉和湖北封城之后,全国各地采取了许多限制性的公共卫生措施,例如旅行限制、隔离、强制戴口罩、禁止多人聚会、暂停公共服务等。这些措施成功地限制了大多数地区疫情的社区传播,使得疫情的传播最多只能发生在家庭之中。我们的模型对大多数地区的高预测能力也表明了在全国各地这些防控措施的一致性和有效性。

3　各种人口流动的风险评估分析

进一步,我们评估从武汉及湖北流出的各种人口数量对全国各地疫情发展的影响。但由于各种流出人口数量之间具有很高的相关性,将这些变量同时引入模型(1)会产生严重的共线性,从而导致模型参数不稳定和不可解释。为解决这一难题,我们采用条件独立性检验(CI)和随机森林(RF)两种非参数统计方法来检验有多个疫情源流入人口对各地疫情影响的大小[1]。这里我们介绍随机森林分析的结果。

决策树和随机森林是解释性较强的非参数模型机器学习方法。RF 模型采用多个决策树的集合与封装以减少单棵树的误差和提高预测的准确性[30,31]。一个简单的随机森林模型由大量的决策树(或回归树)构成,其总体预测值是考虑各个树预测结果的平均值。由于单棵树将产生多维阶跃型函数,因此所有子树的平均值仍然是一个多维阶跃型函数,但也可用于拟合和预测平滑型函数即应用于回归任务。我们的分析采用包含 300 棵回归树的随机森林模型,基于 296 个地级市的样本数据拟合 RF 模型,构建包括各种流动性数据以及本地人口和 GDP 为特征向量作为模型的输入,确诊病例数的对数作为模型的拟合目标。根据包外(Out-Of-Bag,OOB,即没有被决策树选择的数据)估计的结果确定 RF 模型的检验误差,并采用均方误差(MSE)作为评价决策树分支质量的函数。MSE 减少量一般作为回归随机森林中的特征重要性(贡献度)的评价指标。基尼(Gini)重要性是在分类问题中常用于度量随机森林中变量重要性的指标,它是根据指定特征在不同决策树分支上的基尼系数综合计算得到。为便于比较,对所有变量的重要性指标作归一化处理(总和为 1)。

在 1 月 1 日至 24 日期间,有 11 478 484 人次从武汉流向全国 296 个地级市,其中 75.66% 流入湖北省内(除武汉以外),而从湖北其他地区流向全国各地(不包括湖北省)也达 11 379 461 人次。因此,在 1 月 24 日之前从湖北省各地区流入全国其他地区的人口数量同样存在潜在风险。在 Jia 等[1]研究中,我们采用模型(1)和其他模型对武汉流出的人口和湖北其他地区流出的人口进行了比较,结果显示后者对于全国(除湖北以外)的疫情发展没有显著影响。另外,我们还采用随机森林分析的方法对这两种人口流动以及其他因素进行了对比分析,发现武汉流入人口对全国各地疫情的影响是自湖北其他地区流入人口的 12.5 倍。因此,多种分析说明除武汉以外湖北没有形成新的疫情中心,这意味着湖北各地的封城是及时的。

武汉以及湖北其他地市封城后,仍然有一定数量的人口从这些地区流入其他城市。在 1 月 24 日之后的一周里,每天自武汉流入到湖北其他地方的人数为 20 000 左右,流入到湖北省以外的有 1 000 人左右。尽管这些人里面可能包括政府、医疗、后勤服务或其他工作人员,但仍然有必要评估他们对各地疫情的影响。RF 模型的分析显示,在 1 月 25 日至 2 月 6 日期间,从武汉额外流出的人口数量以及从湖北其他地区流出的人口数量对全国各地疫情发展的影响可以忽略不计,其影响分别都低于 1 月 1 日至 24 日期间从湖北(除武汉以外)流入全国各地人口对疫情的影响。

从武汉流向全国各地的人口按照居住属性可以分为两类:一类是武汉市常住人口从武汉去外地,另一类是非武汉市常住人口,即外地去武汉然后流向外地的人口。RF 模型的分析显示,属于武汉居民流出的人口要比去了武汉再返回本地的人口风险大 3.5 倍。这应该是由于武汉居民暴露在新冠疫情环境中的时间和机会要比外地来武汉的人多得多,因此他们更容易成为感染者或带菌者。

最后,我们对于武汉流出人口不同年龄段和性别对于全国各地疫情的影响进行分析。图 4 展示了武汉流出人口不同年龄段与各地确诊人数之间的相关性随时间变化的趋势,其中 18 岁以下这个年龄段(主要为中小学生手机用户)与各地确诊人数的相关性最高,然后是 19～29 岁这个年龄段的年轻人。采用模型(1)对各个年龄段进行分析的结果也同样显示这两个年龄段的参数值最大,即儿童、学生和年轻人的流动对各地疫情发展的驱动作用更显著。为了同时比较不同年龄段武汉流出人口对全国各地确诊人数和死亡人数的影响大小,我们采用 RF 模型来进行分析,其结果见表 1。分析显示,18 岁以下的武汉流出人口对于各地累计确诊人数的影响权重是其他年龄组的 2～3 倍,这一年龄组对于各地累计死亡人数的影响权重也是最大的。因此,武汉流出人口中 18 岁以下的儿童和青少年学生在驱动全国各地疫情扩散中比其他年龄段的人口起着更加重要的作用。目前,有关儿童和青少年在新冠肺炎疫情传播中作用还有待研究[32-34],但相关观点认为病毒释放可能来自于无症状的儿童和青少年,他们会通过学

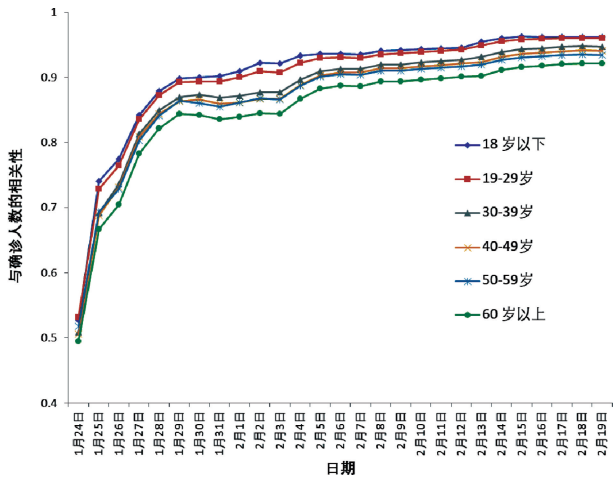


图 4 武汉流出人口不同年龄段与各地疫情的相关性

表 1 不同年龄段对各地疫情影响的随机森林模型分析

变量	累计病例影响（2 月 19 日）	累计死亡影响（2 月 19 日）
18 岁以下	0.323	0.184
19～29 岁	0.141	0.097
30～39 岁	0.110	0.126
40～49 岁	0.151	0.133
50～59 岁	0.104	0.161
60 岁以上	0.129	0.113
GDP	0.023	0.068
人口	0.018	0.119
R^2-Score	0.980	0.941

表 2 不同性别对各地疫情影响的随机森林模型分析

变量	累计病例影响（2 月 19 日）	累计死亡影响（2 月 19 日）
男性	0.453	0.496
女性	0.470	0.456
GDP	0.042	0.032
人口	0.036	0.017
R^2-Score	0.977	0.973

校这一桥梁把病菌带入家庭和社区。虽然儿童和青少年对于新冠肺炎有较低的易感性（中国、意大利、美国都低于 2%）和死亡率，但我们的数据分析显示他们在病菌的传播中起着重要的推动作用。

从相关性分析和随机森林模型分析来看，武汉流出人口中女性对于各地确诊人数的重要性影响要比男性稍微高一些，但男性对于各地死亡人数的重要性影响要比女性更大一些（表 2）。已有的数据和分析都显示，男性比女性感染新冠肺炎后的死亡率要显著的高一些。

4 结论与总结

本项研究利用电信运营商掌握的海量手机用户位移数据，基于疫情源流出人口数量构建了疫情风险的时空发展模型，来检测人们的流动性特征可以在多大程度上捕获新冠肺炎疫情的时空传播规律。与大多数流行病学预测模型不同的是，我们的模型是基于真实的人口流动数据来预测疫情的地域分布和传播趋势。对于新冠肺炎疫情，我们的模型至少提前一周预测了全国范围内感染的病例数量和地理分布，因此，这一方法可以用来监测和预警疫情的早期发展，为制定应急计划和相关决策提供依据。

从武汉流入全国各地的人口数量为各地的疫情风险提供了一个基准。通过建立疫情发展的基准风险时空模型，可用来判断哪些地区的实际疫情显著偏离了它们应该的发展趋势（95%置信区间），从而建立了疫情社区传播风险指数，用以监测各地疫情管控情况。有关部门的决策者可以利用风险指数在疫情爆发初期进行快速、准确的风险评估，以最大限度地控制疫情蔓延。

本文定量研究了我国在疫情应对方面，特别是严格的隔离措施在抑制人口流动和控制疫情蔓延的有效性，分析了各类流动人口的风险大小，并间接评估了各地的应对措施，包括封城、封区、关闭学校、延长假期、人员隔离、限制出行、交通管制等在时间和强度方面对抑制疫情的效果。新冠肺炎疫情能在我国得以控制，得益于各地各部门广泛应用大数据进行疫情风险管控，取得了很好的成效。基于个体层面的手机位置数据来预测疫情风险并定量评价各种管控措施的有效性，为今后的应急管理提供了大数据时代的决策依据。

参 考 文 献

[1] Jia JS, Lu X, Yuan Y, et al. Population flow drives spatio-temporal distribution of COVID-19 in China. Nature, 2020, 582: 389—394.

[2] Barbosa H, Barthelemy M, Ghoshal G, et al. Human mobility: models and applications. Physics Reports, 2018, 734: 1—74.

[3] Brockmann D, Hufnagel L, Geisel T. The scaling laws of human travel. Nature, 2006, 439(7075): 462—465.

[4] Gonzalez MC, Hidalgo CA, Barabasi AL. Understanding individual human mobility patterns. Nature, 2008, 453(7196): 779—782.

[5] Song C, Qu Z, Blumm N, et al. Limits of predictability in human mobility. Science, 2010, 327(5968): 1018—1021.

[6] Deville P, Song C, Eagle N, et al. Scaling identity connects human mobility and social interactions. Proceedings of the National Academy of Sciences, 2016, 113(26): 7047—7052.

[7] Eubank S, Guclu H, Kumar VSA, et al. Modelling disease outbreaks in realistic urban social networks. Nature, 2004, 429(6988): 180—184.

[8] Colizza V, Vespignani A. Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: theory and simulations. Journal of Theoretical Biology, 2008, 251(3): 450—467.

[9] Colizza V, Barrat A, Barthélemy M, et al. The role of the airline transportation network in the prediction and predictability of global epidemics. Proceedings of the National Academy of Sciences, 2006, 103(7): 2015—2020.

[10] Balcan D, Colizza V, Gonçalves B, et al. Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences, 2009, 106(51): 21484—21489.

[11] Brockmann D, Helbing D. The hidden geometry of complex, network-driven contagion phenomena. Science, 2013, 342(6164): 1337—1342.

[12] Chretien JP, Riley S, George DB. Mathematical modeling of the West Africa Ebola epidemic. Elife, 2015, 4: e09186.

[13] Bansal S, Chowell G, Simonsen L, et al. Big data for infectious disease surveillance and modeling. The Journal of Infectious Diseases, 2016, 214(suppl 4): S375—S379.

[14] Pentland A. Reality mining of mobile communications: toward a new deal on data// Dutta S, Mia I eds. The Global Information Technology Report 2008—2009. Geneva: SRO-Kunding, 2009: 75—80.

[15] Giannotti F, Pedreschi D. Mobility, data mining and privacy: geographic knowledge discovery. Germany: Springer Science & Business Media, 2008.

[16] Shaman J, Karspeck A. Forecasting seasonal outbreaks of influenza. Proceedings of the National Academy of Sciences, 2012, 109(50): 20425—20430.

[17] Mossong J, Hens N, Jit M, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Medicine, 2008, 5(3): e74.

[18] Kitsak M, Gallos LK, Havlin S, et al. Identification of influential spreaders in complex networks. Nature Physics, 2010, 6(11): 888—893.

[19] Althouse BM, Scarpino SV, Meyers LA, et al. Enhancing disease surveillance with novel data streams: challenges and opportunities. EPJ Data Science, 2015, 4(1): 1—8.

[20] Reich NG, Brooks LC, Fox SJ, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. Proceedings of the National Academy of Sciences, 2019, 116(8): 3146—3154.

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

[22] Li R, Pei S, Chen B. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). Science, 2020, 368(6490): 489—493.

[23] Chinazzi M, Davis JT, Ajelli M, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science, 2020, 368(6489): 395—400.

[24] Gilbert M, Pullano G, Pinotti F, et al. Preparedness and vulnerability of African countries against importations of COVID-19: a modelling study. The Lancet, 2020, 395(10227): 871—877.

[25] Rader B, Scarpino SV, Nande A, et al. Crowding and the shape of COVID-19 epidemics. Nature Medicine, 2020: 1—6.

[26] Kraemer MUG, Yang CH, Gutierrez B, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. Science, 2020, 368(6490): 493—497.

[27] Lai S, Ruktanonchai NW, Zhou L, et al. Effect of non-pharmaceutical interventions to contain COVID-19 in China. Nature, 2020, 585: 410—413.

[28] Ruktanonchai NW, Floyd JR, Lai S, et al. Assessing the impact of coordinated COVID-19 exit strategies across Europe. Science, 2020, 369(6510): 1465—1470.

[29] Newville M, Stensitzki T, Allen DB, et al. LMFIT: Non-linear least-square minimization and curve-fitting for Python. Astrophysics Source Code Library, Record, 2016: 1606.014.

[30] Grömping U. Variable importance assessment in regression: linear regression versus random forest. The American Statistician, 2009, 63(4): 308—319.

[31] Altmann A, Toloşi L, Sander O, et al. Permutation importance: a corrected feature importance measure. Bioinformatics, 2010, 26(10): 1340—1347.

[32] Snape MD, Viner RM. COVID-19 in children and young people. Science, 2020, 370(6514): 286—288.

[33] Saleem H, Rahman J, Aslam N, et al. Coronavirus disease 2019 (COVID-19) in children: vulnerable or spared? a systematic review. Cureus, 2020, 12(5): e8207.

[34] Swann OV, Holden KA, Turtle L, et al. Clinical characteristics of children and young people admitted to hospital with covid-19 in United Kingdom: prospective multicentre observational cohort study. BMJ, 2020, 370: m3249.

Risk Analysis of COVID-19 Based on Population Flow

Jia Jianmin[1*] Yuan Yun[2] Jia Jayson S[3]

1. *Shenzhen Finance Institute/Shenzhen Institute of Artificial Intelligence and Robotics for Society/ School of Management and Economics, The Chinese University of Hong Kong, Shenzhen* 518172
2. *School of Economics and Management, Tsinghua University, Beijing* 100084
3. *Faculty of Business and Economics, The University of Hong Kong, Hong Kong SAR* 999077.

Abstract This paper reports the research outcomes of our National Natural Science Foundation of China COVID-19 special project. Our primary results were published in Nature (Jia et al. 2020). We also report extensions of this work. First, we report a spatio-temporal benchmark hazard model based on population outflow from Wuhan to other prefectures in mainland China before the lockdown, which successfully predicts and explains the spatio-temporal distribution of COVID-19 across China. Second, we report a risk index for community transmission risk that can be used as a risk detection toolkit assessing epidemic risk for different areas during the early stages of the COVID-19 pandemic. In particular, we estimated transmission risk based on the deviation between confirmed cases and the prediction of the benchmark model. This index can serve as an epidemic warning system to identify and track which regions have a high level of community transmission risk. Finally, we present statistical models and machine learning approaches that can assess heterogeneities in risk sources of population outflows and test the relative contributions of different risk sources on the spread of COVID-19; for example, we estimate the relative contribution of Wuhan versus non-Wuhan residents, men versus women, and different age groups on COVID-19 transmission risk.

Keywords population flow; COVID-19; epidemic transmission; spatio-temporal hazard model; risk analysis

(责任编辑 姜钧译)

* Corresponding Author, Email: jmjia@cuhk.edu.cn