建模探索

人口迁入与新增确诊数的趋势关系及因果量化分析

廖 可[1]，胡云鹤[1,2]，侯馨翔[1]，吴凌霄[1]，王 杰[1]，张 一[3]，戴彧虹[4]，杨周旺[1]

(1. 中国科学技术大学，安徽 合肥 230026；2. 华东师范大学，上海 200241；
3. 北京大数据研究院，北京 100871；4. 中国科学院 数学与系统科学研究院，北京 100190)

摘 要：目前，很多地区新冠肺炎疫情已得到缓解，复工、复产已被多地政府部门提上日程. 2 月 10 日前后，全国各地返城复工人数增多，2 月 14 日开始，广东、河南等地新增病例数出现了明显反弹，人口跨地区迁徙使疫情防控更加困难. 目前全国返工、返校需求还远未得到满足，需要通过数据分析，对“返城复工”的风险进行评估. 通过观察数据可以发现人口迁徙与新增确诊病例数有很强的正相关性，因此由“格兰杰因果检验”确定了人口迁徙与新增确诊病例数有显著的因果关系.

关键词：新型冠状病毒；感染者；确诊病例；人口迁徙；因果分析

中图分类号：O29 **文献标志码**：A **文章编号**：2095-3070(2020)01-0023-06

0 引言

自新型冠状病毒疫情爆发以来，举国上下皆倾尽全力与之对抗. 目前，很多地区的疫情已经得到缓解. 然而，疫情的缓解是以社会经济发展为代价，从这一角度考虑，允许人们返回工作场所复工，不是“如果”，而是“何时”和“如何”的问题. 从 2 月份开始，除湖北以外地区的新增确诊病例已经从“以有湖北旅居史的病例为主”逐步发展为“以无湖北旅居史的病例为主”，肺炎防控进入新阶段. 紧接着，在 2 月 10 日前后，全国各地返城复工人数增多，2 月 14 日开始，广东、河南等地新增病例数出现了明显反弹. 这说明当前的肺炎防控形势依旧严峻，需要通过数据分析模型清晰地给出“返城复工”的时间选择和风险. 通过观察，发现人口迁徙与新增确诊病例数有很强的正相关性，进而通过“格兰杰因果检验”确定了人口迁徙与新增确诊病例数有显著的因果关系，这为开展复工、复产的策略优化研究奠定了基础.

1 疫情数据初步分析

疫情爆发以来，各地政府部门通过多种渠道积极发布疫情信息，为我们提供了有力的数据支撑. 为了明确，在此先说明一代、二代与三代病例的定义如下：

一代病例是指来自疫情发生地的输入型病例，即有湖北旅行史或居住史的新型冠状病毒感染者；

二代病例是指受输入型病例感染的病例，即无湖北旅行史或居住史，但有湖北人员接触史的新型冠状病毒感染者；

三代病例是指受二代病例感染，无湖北旅居史，又无湖北人员接触史的新型冠状病毒感染者.

1.1 一代与非一代病例情况对比

可以看出，区分一代与非一代病例的标志在于有无湖北旅居史. 通过收集病例的旅居史信息，分

收稿日期：2019-02-20
基金项目：国家自然科学基金(71950011，11871447，71991464/71991460)；国家重点研发计划课题(2018AAA0101001)
通讯作者：杨周旺，E-mail：yangzw@ustc.edu.cn

离出各地区每日新增的一代、非一代病例数量，发现多数观察地区的新增确诊病例已经从“一代病例为主”阶段发展到“非一代病例为主”阶段. 以浙江省重点城市杭州市和温州市为例(图 1 和图 2)，两个阶段的交接点多是在武汉宣布“封城”(1 月 23 日)以后的一周(1 月 30 日)左右.

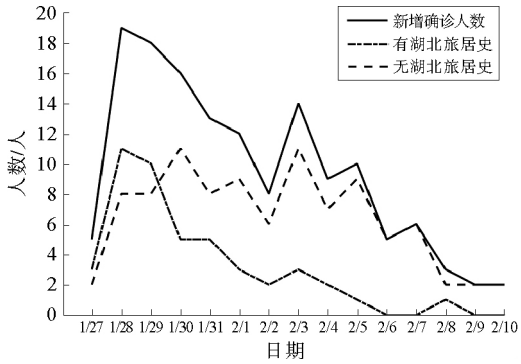

图 1　杭州市有无湖北旅居史病例数对比图

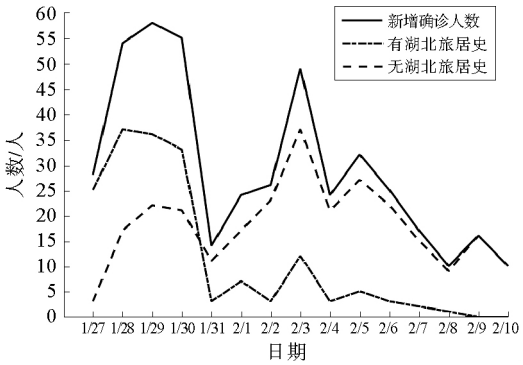

图 2　温州市有无湖北旅居史病例数对比图

1.2　深圳市病例数分析

同时，仍然存在少数城市，例如深圳市的新增确诊病例以一代病例为主(图 3). 虽然不排除这些有湖北旅居史的感染者是在离开湖北后才被传染的可能性，但也表明了超过两周潜伏期的病毒可能不在少数.

通过与百度迁徙的人口迁入指数(反映日人口迁入数量)进行对比，如图 4 所示，可以发现，由于常住深圳的湖北籍人口在节后返回深圳，深圳市新增确诊病例数(主要为一代病例)在 1 月 25 日之后经历了 2 次高峰(分别为 1 月 31 日和 2 月 3 日). 因此可以推测，正是由于深圳市人口迁徙规模如此之大，才造成了一代病例数所占比例长期居高不下.

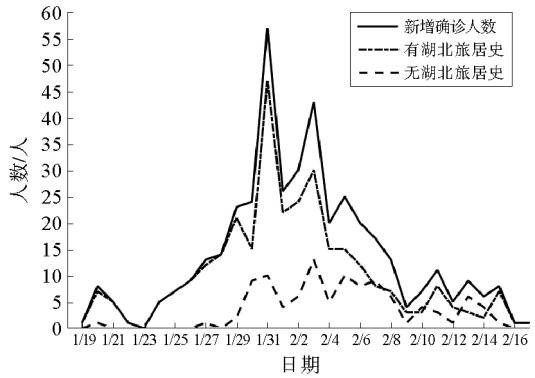

图 3　深圳市有无湖北旅居史病例数对比图

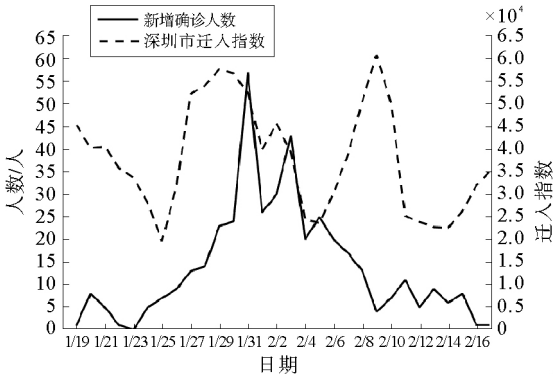

图 4　深圳市新增确诊数与迁入指数对比图

进一步观察发现，还有很多地区迁入指数与新增确诊病例数两个时间序列呈现非常显著的正相关性. 将各地的新增确诊病例数回溯 2～6 天(新增病例曲线向前平移 2～6 天)，呈现与百度迁徙的人口迁入指数曲线较好的趋势一致性. 下一节利用因果量化分析进一步确认人口迁徙与新增确诊病例数的因果关系.

2　人口迁入与新增病例数的趋势关系及因果量化分析

收集了 1 月 20 日至 2 月 10 日期间北京市、上海市、广东省、浙江省和深圳市的相关数据，通过观察分析这些省市人口迁入指数与回溯若干天后的新增确诊病例数，可以发现两个时间序列呈现非常明显的正相关性，且大体趋势高度一致(其中北京市、上海市、浙江省如图 5—图 7 所示). 如果能够找到一种方法，来证明迁入指数与新增病例数确实具有一定程度的因果关系，这将有助于未来时刻新增确

诊数的预报，以及返城复工期间疫情防控政策的制定.

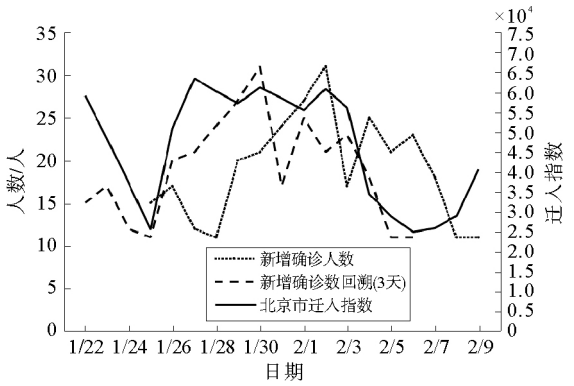

图 5　北京市新增确诊数与迁入指数对比图
（截取 1 月 25 日至 2 月 9 日数据）

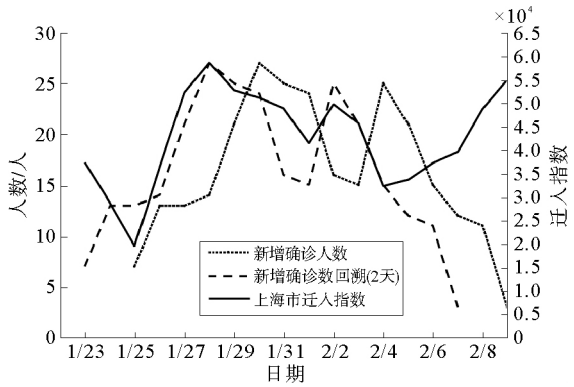

图 6　上海市新增确诊数与迁入指数对比图
（截取 1 月 25 日至 2 月 9 日数据）

2.1　格兰杰因果分析简介

“格兰杰因果关系检验”(Granger causality test)是一种衡量时间序列之间相互影响的方法，由 2003 年诺贝尔经济学奖得主克莱夫·格兰杰(Clive W. J. Granger)提出[1]，早期多用于计量经济学中的变量预测. 最近十几年，这一方法还被广泛用于经济学以外的其他领域，包括气象科学、神经科学等.

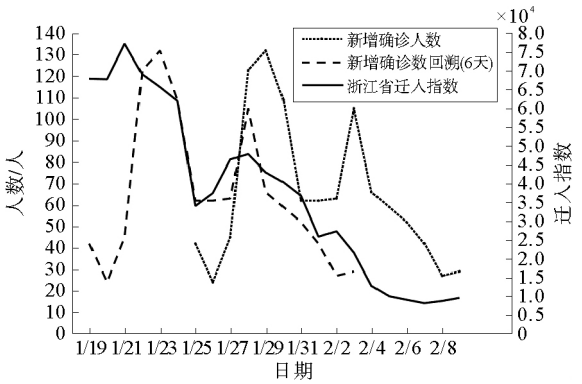

图 7　浙江省新增确诊数与迁入指数对比
（截取 1 月 25 日至 2 月 9 日数据）

为了说明格兰杰因果关系，首先假设有一个时间序列 y_1，它由不同时刻获得的 n 个样本数据(y_{11}，y_{12}，…，y_{1n})组成，再假设另一个时间序列 y_2，类似地，它由(y_{21}，y_{22}，…，y_{2n})组成. 然后，先利用 y_2 的过去预测 y_2 的未来，得到一个预测误差 δ_2；再利用 y_1 和 y_2 共同的过去预测 y_2 的未来，得到另一个预测误差 δ_1. 如果 $\delta_1<\delta_2$，即 y_1 和 y_2 的联合预测误差小于 y_2 自身预测误差，那么这说明 y_1 有助于预测和解释 y_2，此时称 y_1 对 y_2 有格兰杰因果关系[2].

至于预测具体是如何实现的，需要引入自回归和联合回归的概念. 下面两个公式分别展示了这两种回归的具体原理[3]：

$$\text{自回归：}\ y_{2t}=\alpha_0+\sum_{k=1}^{m}\alpha_k y_{2,t-k}+\varepsilon_t, \tag{1}$$

$$\text{联合回归：}\ y_{2t}=\beta_0+\sum_{k=1}^{m}\alpha_k y_{2,t-k}+\sum_{k=1}^{m}\beta_k y_{1,t-k}+\tilde{\varepsilon}_t. \tag{2}$$

其中，参数 α_k，β_k ($k=0$，1，…，m)可以通过最小二乘法获得相应估计. 如果忽略误差项 ε_t 和 $\tilde{\varepsilon}_t$，首先将具体数据代入式(1)，便可得到自回归情形下序列 y_2 在 t 时刻的预测；再将具体数据代入式(2)，便能得到联合回归情形下序列 y_2 在 t 时刻的预测. 预测值与真实值的差异即是预测误差，上文提到的 δ_2 和 δ_1 分别是这两种情形下的均方预测误差.

在得到 δ_2 和 δ_1 的值后，除了比较二者大小，还需要判断该大小差异是否显著，因此采用统计假设检验中常用的 F 检验. 检验的原假设是 y_1 不是 y_2 的格兰杰原因，显著性水平取为 0.05. 如果检验 p 值小于 0.05，则拒绝原假设，即可以认为存在显著的格兰杰因果关系，且 y_1 是 y_2 的因[4].

至此，大致了解了格兰杰因果分析的原理，下面将进一步介绍所使用的建模方法.

2.2　疫情建模方法

将人口迁入指数时间序列记为 y_1，将每日新增确诊病例数时间序列记为 y_2，将 m 阶向量自回归模型记为 VAR(m)，目标是利用向量自回归模型来验证人口迁入指数是每日新增确诊病例数的格兰杰原因. VAR(m)模型的结构如下所示[5]：

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = c + \theta t + \begin{bmatrix} a_{11,1} & a_{12,1} \\ a_{21,1} & a_{22,1} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \cdots + \begin{bmatrix} a_{11,m} & a_{12,m} \\ a_{21,m} & a_{22,m} \end{bmatrix} \begin{bmatrix} y_{1,t-m} \\ y_{2,t-m} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}. \tag{3}$$

模型阶数 m 限定了模型在对未来进行预测时所能用到的最近的历史数据，例如 m 取 2 时，模型最多只能使用两天前的历史数据. 受制于待估参数的数量，阶数 m 在目前样本量情况下最大能取到 6. 考虑到迁入指数对新增病例数的影响可能具有一定的时间滞后性，而阶数 m 的差异正好体现了这种影响，从新增病例数序列的角度来看，希望得出较为合理的回溯时间(天). 因此，建立不同阶数的 VAR(m)模型，并进行格兰杰因果关系检验. 由于检验的 p 值在一定意义上体现了 y_1 对改善模型预测效果的显著性程度，p 值越小，在设定模型阶数下的 y_1 作用越显著，所以选取 p 值最小的模型所对应的阶数作为理想的回溯天数.

在一般的格兰杰因果检验中要求序列为平稳时间序列，简单来说就是序列不应该有明显的变化趋势，不同时间段数据的方差也不应该有明显的差异. 因此首先进行 KPSS 非参数平稳性检验，检验结果表明所选用省市对应的两条时间序列都满足要求.

2.3　量化分析结果

量化分析所使用的原始数据如表 1 所示.

表 1　各地区新增确诊数与迁入指数表

地区	北京市		上海市		广东省		浙江省		深圳市	
日期	新增病例数	迁入指数	新增病例数	迁入指数	新增病例数	迁入指数	新增病例数	迁入指数	新增病例数	迁入指数
1 月 25 日	15	25 494	7	19 504	20	43 383	42	33 996	7	19 644
1 月 26 日	17	50 772	13	36 313	48	68 269	24	37 441	9	31 700
1 月 27 日	12	63 442	13	52 192	42	111 046	45	46 423	13	52 260
1 月 28 日	11	60 129	14	58 600	53	123 841	123	47 847	14	54 069
…	…	…	…	…	…	…	…	…	…	…
2 月 6 日	23	24 902	15	37 218	74	74 301	52	9 017	20	30 307
2 月 7 日	18	25 874	12	39 593	57	95 650	42	8 120	17	38 986
2 月 8 日	11	28 899	11	48 733	45	128 113	27	8 702	13	50 728
2 月 9 日	11	40 554	3	54 914	31	140 240	29	9 524	4	60 909

所选取的 5 个省市的格兰杰因果关系检验结果如表 2 所示.

表 2　量化分析结果表

地区	北京市	上海市	广东省	浙江省	深圳市
结论	显著	显著	显著	显著	显著
F	8.302 5	7.466 4	4.676 8	36.338 0	5.692 4
p-value	0.004 6	0.014 8	0.042 6	0.001 9	0.042 0
回溯期/天	3	2	3	6	4

可以看到，通过对以上 5 个省市进行格兰杰因果检验，结果表明这些省市都可以得出人口迁入指数是新增病例数的格兰杰原因，将联系实际进行分析解释，来更好地印证提出的观点：人口迁入指数的变动在一定程度上引起了若干天后新增病例数的同方向变动[6]. 从实际情形来看，人口迁徙一方面可能会带来感染者，另一方面还会增加易感人群数量. 因此，当人口迁入规模扩大时，当地确诊病例数

很可能在几天后随之增加；当人口迁入规模缩减时，当地确诊病例数则很可能在几天后随之减少[7-8].

根据各省份、城市的量化分析结果，归纳总结如下：

第一，各省份、城市的新增确诊数在回溯一定时间后，能够与其迁入指数趋势进行较好的匹配，其中上海市、浙江省的新增确诊数回溯后都能长期与迁入指数在趋势上保持高度吻合.

第二，各省份、城市最优回溯时间有所不同，经分析主要原因有两点：

1)各省市医疗水平、政策上的不同，导致确诊病人的效率不同，例如上海市最优回溯时间为 2 天，在所有省市中最短，这得益于其优异的医疗水平和确诊能力；

2)各省市人口流入结构不同，例如浙江省以春节前的返乡人员为主，广东省则有大量节后复工人员. 因此浙江省的确诊患者一般在潜伏期早期回到浙江，回溯时间长达 6 天，广东省的确诊患者一般在潜伏期晚期或已有症状时回到广东省，回溯时间相对较短，大致是 3 天.

3　返城复工潮对疫情防控的影响和基本建议

新冠肺炎疫情爆发以来，已对国家经济造成巨大的破坏，各行各业急需恢复生产. 各地区政府部门需要合理安排企业复工、复产时间表，严格制定工作生产过程中的防疫措施，循序渐进地恢复当地经济，防止疫情二次爆发. 未来一个月内，出现全国范围内的大规模人口迁徙几乎是必然的，由人口迁徙与新增病例数的量化分析，可以知道人口迁徙与疫情发展情况息息相关. 于是，需要在人口大量涌入各个省市时做好以下准备：

首先，不可因为疫情一时缓解，便放松在火车站、客运站、机场等地的检测力度，要坚持对所有疑似人员集中隔离，尽量降低人口迁徙带来的疫情二次爆发风险. 同时，要对不同情况的迁入人口进行有针对性的限制，降低迁入人口成为病毒传播者的可能性.

其次，有计划地指导各个企业恢复正常运营，要严格执行员工防疫措施，尽量避免聚集性病例的发生，严防大规模传染. 各个企业需要每日统计所有员工的健康状况，并及时上报有疑似症状的员工. 同时，对有条件的企业，建议尽量线上办公，降低接触的可能性.

最后，坚持对街道小区、产业园区、校园、办公楼宇、商店等实行封闭式管理，尤其是社区，需要完善外来人员的健康信息登记，及时将疑似人员上报疾病控制中心.

为应对接下来的返工、返校潮，需要做好以上准备，这样可以在最大程度上降低人口迁徙导致疫情扩散的可能性，控制疫情的发展.

致谢：感谢北京大数据研究院、中国科学院数学与系统科学研究院、西安交通大学金融优化团队参与讨论，以及“柚子优化”微信公众号团队和国科优化数据团队提供宝贵意见和多方面的支持.

参考文献

[1]Granger C W J. Testing for causality: A personal viewpoint[J]. Journal of Economic Dynamics and Control, 1980, 2(1): 329-352.

[2]CSDN. 漫谈格兰杰因果关系(Granger Causality)——第一章 野火烧不尽，春风吹又生[EB/OL]. (2017-08-14). https://blog.csdn.net/ggabcda/article/details/77165104.

[3]Wang X. A Granger causality test of the causal relationship between the number of editorial board members and the scientific output of universities in the field of chemistry[J]. Current Science, 2017, 116(1): 35-39.

[4]李秋玲，王智文，张灿龙，等. 基于 VAR 模型的 CPI 与 PPI 因果关系检验[J/OL]. 广西科技大学学报，2020(01): 104-110. https://doi.org/10.16375/j.cnki.cn45-1395/t.2020.01.016.

[5]Bose E, Hravnak M, Sereika S M. Vector autoregressive models and Granger causality in time series analysis in nursing research: Dynamic changes among vital signs prior to cardiorespiratory instability events as an example[J]. Nursing Research, 2017, 66(1): 12.

[6]丁启燕，杨振，周晴雨. 中国艾滋病疫情分布变化与人口流动性研究[J]. 热带地理，2017, 37(04): 538-546.

[7]武继磊，王劲峰，孟斌，等. 2003 年北京市 SARS 疫情空间相关性分析[J]. 浙江大学学报：农业与生命科学版，

2005(01): 100-104.
[8]曹志冬, 王劲峰, 高一鸽, 等. 广州 SARS 流行过程的空间模式与分异特征[J]. 地理研究, 2008(05): 1139-1149+1226.

The Quantitative Analysis of Causality between Population Migration and the Number of Newly Confirmed Cases

LIAO Ke[1], HU Yunhe[1,2], HOU Xinxiang[1], WU Lingxiao[1], WANG Jie[1], ZHANG Yi[3], DAI Yuhong[4], YANG Zhouwang[1]

(1. University of Science and Technology of China, Hefei, Anhui 230026, China;
2. East China Normal University, Shanghai 200241, China;
3. Beijing Institute of Big Data Research, Beijing 100871, China;
4. Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: At present, the epidemic situation of COVID-19 has been alleviated in many areas, so the resumption of work and production have been put on the agenda by many local governments. Around February 10th, the number of workers back to work across the country increased greatly. And since February 14th, the number of newly confirmed cases of COVID-19 in Guangdong, Henan and other provinces has rebounded significantly. The migration of population across regions has made it more difficult to prevent and control the epidemic. Currently, the needs for people back to work and school across the country are far from being met, and data analysis can be used to evaluate the risk of "return to the city and back to work". It can be found that there was a strong positive correlation between population migration and the number of newly confirmed cases of COVID-19 from existing data, and the result of "Granger Causality Test" confirmed that there was a significant causal relationship between them.

Key words: COVID-19; infected people; confirmed cases of COVID-19; population migration; Granger causality test

作者简介

廖　可(1996—), 男, 中国科学技术大学大数据学院硕士研究生, 主要研究方向为经济大数据建模.

胡云鹤(1998—), 男, 华东师范大学统计学院本科生, 主要研究方向为经济大数据建模.

侯馨翔(1998—), 女, 中国科学技术大学数学科学学院本科生, 主要研究方向为经济大数据建模.

吴凌霄(1998—), 男, 中国科学技术大学数学科学学院硕士研究生, 主要研究方向为应用数学、机器学习.

王　杰(1997—), 男, 中国科学技术大学大数据学院硕士研究生, 主要研究方向为经济大数据建模.

张　一(1986—), 男, 北京大数据研究院研究员, 主要研究方向为经济大数据分析.

戴彧虹(1971—), 男, 中国科学院数学与系统科学研究院研究员, 博士生导师, 主要研究领域为非线性优化计算方法、理论以及在不同领域中的应用.

杨周旺(1974—), 男, 中国科学技术大学数学科学学院教授, 博士生导师, 主要研究领域为“应用数学”, 即综合运用计算科学、统计、最优化等理论为解决相关问题建立新的数学模型, 发展新方法.