面向新冠疫情的数据可视化分析与模拟预测

陈宝权[1] 史明锰[2] 蒋鸿达[1] 等
[1] 北京大学
[2] 山东大学

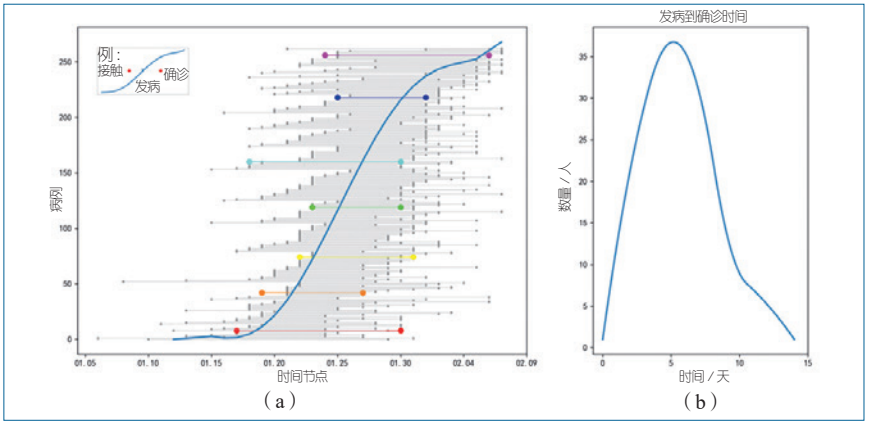
关键词：新型冠状病毒 病毒传播动力学 可视分析

引言

2019年末，由新型冠状病毒引起的呼吸系统疾病（COVID-19）在湖北武汉暴发，并迅速蔓延到中国其他地区和一些国家。中国启动了突发公共卫生事件一级响应，世界卫生组织（WHO）将新型冠状病毒疫情列为国际关注的突发公共卫生事件。新型冠状病毒肺炎（以下简称“新冠肺炎”）的流行不仅危及到人民群众的生命健康，也带来了重大的经济损失。

新冠肺炎的防控阻击战是一项复杂的系统工程，全球各领域的科学家们从不同的侧面对病毒展开科研攻关。其中传染病动力学研究是其中的一大难点。以2020年1月31日《柳叶刀》（*The Lancet*）发表的中国香港科学家的工作[1]为例，许多工作都采用了经典的SEIR模型进行建模，但由于SEIR模型对人群的分类过于简单，理想化的传播过程也难以表达现实环境对传播的影响，导致预测结果存在较大偏差。

为了解决这些难题，我们希望通过科学可视化的方法，从已有数据中挖掘出疫情传播的潜在特点，并根据其特点建立更科学的传染病动力学模型，定性并定量地分析疾病流行的原因和规律，对疫情传播的过程进行更科学的模拟和预测，揭示其规律和模式。



（a） （b）

图1 （a）病例接触、发病、确诊时间可视化；（b）发病时间统计

疫情传播特点

病毒基本传播性质

新冠肺炎病毒携带者是此次疫情最主要的传染

源，潜伏期患者、隐性感染者（即无症状感染者）和恢复期患者也有一定的传染性。为了对新冠肺炎病毒潜伏期等基本性质进行研究，实验组通过手动标注深圳市的200余起病例进行可视化，观察疾病感染、发病、确诊三个阶段的时间关联性，发现疾病的部分传播特点。

在图1（a）的可视化中，以时间为横轴，通过将不同的病例以条状散点的方式平铺在时间轴上展示不同病例的时间分布情况。蓝色分界线表示发病的时间，左侧点表示此人接触其他患者的时间，右侧点表示此人被确诊的时间。

通过病例可视化和发病时间统计（见图1（b）），可以发现大部分新冠肺炎病例的潜伏期在5~7天，最长可达15天。这为我们深入研究新冠肺炎病毒的传播特点奠定了基础。基于这样的潜伏期事实，我们开始扩大范围，对全国的传播情况进行可视化。

全国范围传播分析

扩散性 为了更直观地从整体上对传播过程进行分析，本文首先重现了病毒在全国范围内传播的演变过程。我们使用国家及各省市地区卫健委公布的地级市每日确诊数据[2]，通过热度图的方式对病毒传播进行时间维度上的可视化。热度图的可视化方式可以将离散的地理坐标点通过区域的形式连接起来，区域的传播往往更能发现规律。如图2所示，疫情的传播首先主要以武汉为中心向周围扩散，然后传播至其他人口密集的中心区域，如北京、上海、广州等地，然后再以这些区域为中心进行二次传播。

阶段性 在了解总体趋势之后，我们通过引入其他变量的方式对传播过程进行了进一步分析。根据流行病学调查，呼吸道疾病的传播离不开人口的流动，因此我们引入人口流动数据，继续通过可视化的方式，对两者之间的相关性进行直观展示，并对传播与人口流动之间的相关性进行量化评估。

春运期间，全国范围人口流动数据趋于稳定，我们可以将其视为固定变量。根据对病毒本身特征的研究，我们假设病毒的潜伏期为7天。因此将武汉封城前7天（至1月31日）的累积数据，及7天之后的累积数据进行对比来展示两个阶段的差异。

人口流动数据使用了百度慧眼所提供的公开数据[3]，提供了春运期间各省份之间的人口流动情况。为了更好地进行对比，本文也将两个阶段的确诊数据以省份为单位进行地图级别的可视化。中国大陆各省份的颜色，反映了该省的确诊人数及来自武汉市的输入人流量。同时为了对不同级别的数据进行对比，在可视化之前首先进行了归一化处理。

对比图3（a）和（b）可以看到，在疫情发展的第一阶段，各省市感染总人数与春运期间由武汉市输入的人流量呈现强相关性；但随着时间的推移，各省市感染人数比例逐渐发生变化（图3（c））。从中可以推断出，第二阶段的疫情发展主要以内部传播为主。为此，本文在建立传染病动力学模型时，对这两阶段分别采用不同的模型参数，认为第一阶段各省市感染者主要通过外部输入。由于从1月23日起武汉采取封城措施，各省市也都开始限制人口流动，所以第二阶段则以第一阶段结束作为初始化值，对疫情传播过程进行数值化分析。

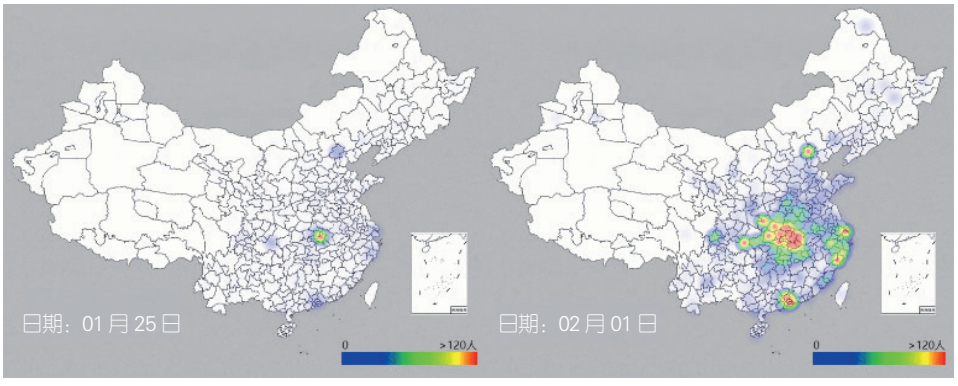


图2 各省市疫情传播热度图（截至1月25日和2月1日的各省市感染总人数）

区域性 在传播的第二阶段，不同的省份产生了明显的省内传播差异，即传播

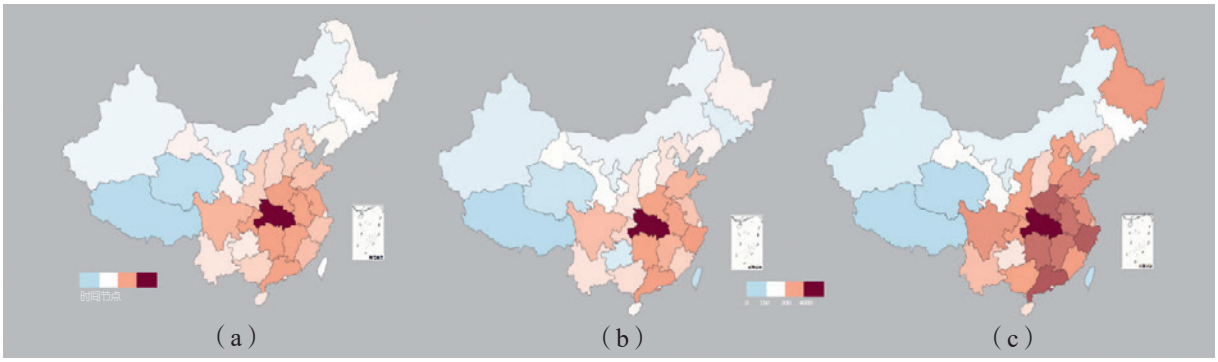


图3 （a）由武汉市流向各省市的输入人流量；（b）1月31日各省市确诊感染总人数；（c）2月9日各省市确诊感染总人数

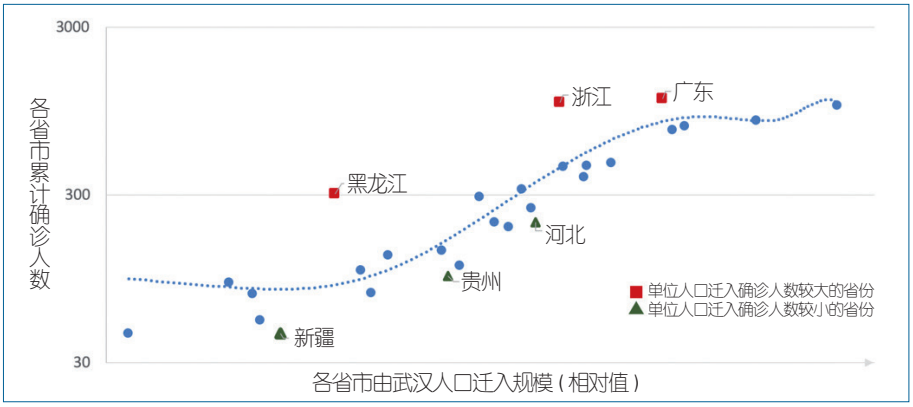


图4 湖北以外省市的武汉人口流入规模（相对值）与其确诊感染人数

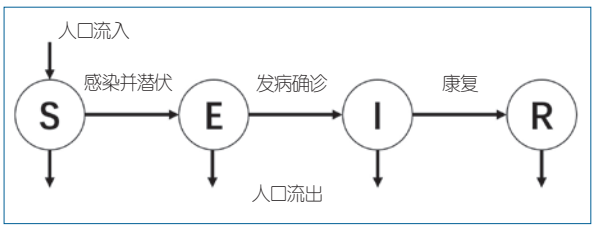


图5 SEIR传染病动力模型

的分区域性。这与各省市人口密度、发展水平等客观条件密切相关。我们通过散点图的方式将不同区域的特征可视化出来。

如图4所示，我们以武汉迁入的人口规模为横轴，所对应省份的累计确诊数量为纵轴，同时使用最小二乘法拟合了对应曲线。这样的可视化主要关注两方面信息：（1）坐标轴的方向所描述的数据的相对规模，（2）与拟合曲线之间的偏差值所代表的数据中的异常。图中的异常点代表了疫情传播比较特殊的几个省份，如黑龙江、浙江等。

综上，通过可视分析可以发现，此次疫情传播具有典型的点状扩散、分阶段、分区域传播的特点。基于此，在用模型对疫情传播模拟的过程中，本文对不同阶段、不同地区需要分别进行独立的参数估计，以保证模拟的合理性。

疫情传播模拟

经典的SEIR模型将人群分为易感人群（Susceptible，S）、已被感染但无症状处于潜伏期的人群（Exposed，E）、已表现出症状但未被隔离的患病人群（Infectious，I）、康复人群（Recovered，R）四类（模型把死亡人数也归到R中），并假设他们之间按一定概率转移，其状态转移如图5所示。

该模型所涉及的参数主要为：可再生数R0、平均潜伏期时间DE和平均收治时间DI。其中，后两种参数均可直接从官方发布的数据中获得，而难点在R0，即一名被感染者平均每天传染到的人数，数值较难准确估计。文献[1]使用了2019年12月31日至2020年1月28日的感染人数数据，并根据境

外（除香港）受感染人数及国际航班从武汉出境人数反推得出 R0 为 2.68，采用如上模型推算出，截至 1 月 25 日，武汉地区受感染人数约为 75815 人，预测疫情的拐点将在 5 月到来，并得出“封城手段的采取对加快疫情缓解的作用不显著”的判断。

我们观察，该项工作的模型与参数选取存在不合理性，主要体现在：

1. 境外确诊数据样本量较小，且使用飞机这一交通工具的人群在总人口中并非均匀分布，据此假设泊松过程来估计 R0 偏差较大。

2. 更重要的是，考虑到政府防控措施的实施与升级，R0 的取值不应设为定值。尽管论文中假设人群戴口罩可以使 R0 减半，并进行了一定的讨论，但这样的设置依然较为粗糙。

3. 封城作为非常严厉的防控手段执行得非常彻底，社区隔离以及疑似病人的隔离措施等控疫手段必须在模型设计与参数设置中进行有效考虑。

C-SEIR 模型

针对 SEIR 模型存在的不足，本文采用 C-SEIR 模型 [4] 对疫情进行模拟分析。 相比于 SEIR 模型，C-SEIR 模型主要有以下两点改进：

1. 考虑政府的隔离措施，将人群进一步划分出隔离患者和未隔离患者，隔离患者不具备病毒传播能力。

2. 考虑政府措施的加强和群众防护意识的上升，病毒的 R0 值应该随时间变化，而不是一个固定值，因此通过真实数据拟合出病毒的传染率曲线代替 R0。

针对第一点，C-SEIR 在 SEIR 的四类人群基础上增加了两类新的人群：被隔离疑似感染人群（P）和已确诊并被隔离的患病人群（Q）。注意在 P 类中的人包括新冠肺炎患者，也包括了症状相似但未感染新冠病毒的人群。我们可以假设被隔离的人群不具备向外传染病毒的能力（实际上被隔离的病人有一定概率感染医护人员，但考虑到被感染的医护人员占总体被感染人群的比例很小，可以忽略不计），即病毒的传染能力只与 I 和 E 有关。

图 6 描述了 C-SEIR 传染病动力学模型的框架。

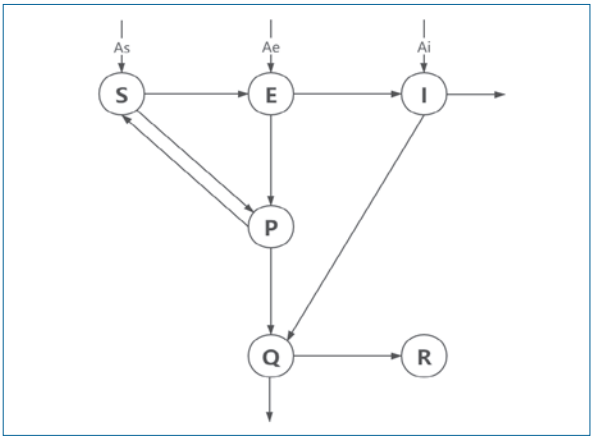


图 6　C-SEIR 传染病动力模型

与 SEIR 模型不同的是，一部分无症状感染者在接受核酸检测后将被妥善隔离，从而进入 P 类，同时感染者的密切接触者以及症状类似的普通流感患者经过排查后也将接受医学观察而进入 P 类，观察若干天后若检测为阴性，则回到 S 当中，否则将被继续隔离接受治疗；已确诊并收治的患病人群 Q 则可能来源于正在接受医学观察的人群 P，或感染发病而未被隔离的人群 I。

从而我们可以定义 C-SEIR 模型的转移方程：

$$
\begin{aligned}
\frac{dS}{dt} &= A_s - \beta(I + kE + z) - d_{sp}Q + d_{ps}P \\
\frac{dE}{dt} &= A_e + \beta(I + kE + z) - d_{ei}E - d_{ep}E \\
\frac{dI}{dt} &= A_i + d_{ei}E - d_{iq}I - \delta I \\
\frac{dP}{dt} &= d_{ep}E + d_{sp}Q - d_{ps}P - d_{pq}P \\
\frac{dQ}{dt} &= d_{pq}P + d_{iq}I - \delta Q - \gamma Q \\
\frac{dR}{dt} &= \gamma Q
\end{aligned}
\tag{1}
$$

下面对方程中出现的参数进行说明。A_s、A_e 和 A_i 表示对应状态的人群从外部的净输入。对于 S 的方程，β 表示每单位时间病毒从携带者转移到疑似患者的概率，k 表示病毒潜伏期相对于发病期的严重程度，z 为初始的感染病人人数，d_{sp} 表示从 S 到 P 的传递速率，但是考虑到 P 远比 S 小，直接使用 $d_{sp}S$ 来衡量会导致病态矩阵的出现，所以我们实际

上使用 $d_{sp}Q$ 替换该部分；$d_{ps}P$ 是从 P 到 S 的传播速率，这表示我们假设 P 中有常数比例的未感染人群。其他方程中的参数的定义和 S 中的 d_{ps} 的定义类似，表示从某一个状态到另一个状态转移的概率。此外，δ 表示未被治疗患者的死亡率，γ 表示受治疗患者的死亡率，与 SEIR 模型不同的是我们不会把死亡人数计入治愈人数之中。

针对之前提到的第二点改进，在C-SEIR模型中，病毒的致病性和人群的防控措施都会随时间发生变化，这就导致参数 β 在模拟的过程中应为一个含时函数，在 t 时刻，新感染的人群数量可以用公式

$$F(t) = \beta(t)\big(I(t) + kE(t)\big) \tag{2}$$

表示。为了得到 β 的具体形式，首先通过政府公布的数据估计 F、I、E 在 t 时刻的值，通过公式得到 β 的近似值。具体方法如下：假设平均的潜伏期是 n_e，病人从出现症状到隔离治疗平均花费的时间为 n_{iq}，官方通报的在 t 时刻新增加的确诊病例为 $\hat{F}(t)$，则有：

（1）$F(t) = \hat{F}\big(t + n_e + n_{iq}\big)$，表示新增的被感染者会在 n_e+n_{iq} 天后被确诊再被官方通报。

（2）$E(t) = \sum_{j=n_{iq}}^{n_e+n_{iq}-1} \hat{F}(t+j)$，表示如果病人在时间 [$t$-$n_{iq}$,$t$-1] 出现症状，它会在当前时间 t 处在 I 状态。

（3）$I(t) = \sum_{j=0}^{n_{iq}-1} \hat{F}(t+j)$，如果病人在时间 [$t$-$n_e$,$t$-1] 进入潜伏期，则在时刻 t，它依然处于潜伏状态 E。

从而，我们可以得到 β 的估计值：

$$\hat{\beta}(t) = \frac{\hat{F}\big(t + n_e + n_{iq}\big)}{\sum_{j=0}^{n_{iq}-1} \hat{F}(t+j) + k\sum_{j=n_{iq}}^{n_e+n_{iq}-1} \hat{F}(t+j)}. \tag{3}$$

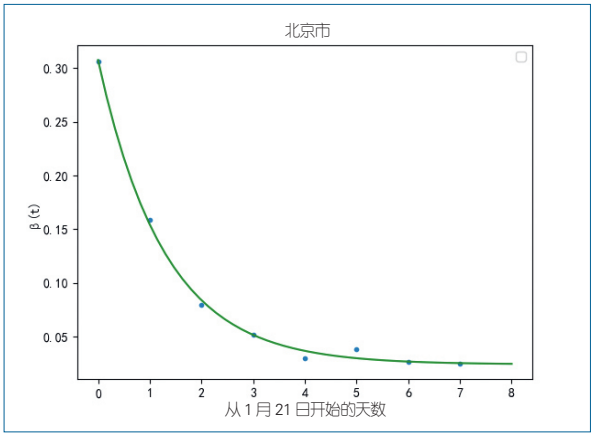


图 7　北京市的感染率变化拟合曲线

图 7 展示了利用这一近似方程和北京市卫健委每天公布的确诊人数得到的北京市的疫情初期的 β 曲线，这里使用了指数函数来拟合得到的散点值，可以发现数据拟合得非常好，这反映了模型的合理性。

模型中的其他参数的估计则可以通过经验性的方法。首先通过公开的信息估计出病毒的平均潜伏期 n_e，病人从发病到被隔离治疗中间花费的时间 n_{iq}，平均被隔离观察的时间 n_p，平均被隔离治疗的时间 n_q，病毒的死亡率 α，然后利用与 SEIR 模型相同的方法，确定 S、E、I、R 四个状态之间的转移概率：假设状态转移过程为泊松过程，平均转移的概率就为 1 除以处在当前状态平均的周期长度。对于 SEIR 状态之外的状态转移概率，如果被感染的病人中被隔离的比例是 r_e，潜伏期是 n_e，可以得到 $d_{ei} = \frac{1-r_e}{n_e}$，$d_{ep} = \frac{r_e}{n_e}$。

类似地，对于 d_{sp}、d_{ps} 和 d_{pq}，首先从政府报告中估计 P 中非 COVID-19 的病人比例 r_s 以及每新增一个确诊病人会新增多少疑似病人 a，并假设平均治疗时间为 n_q 以及平均观察时间为 n_p，则有 $d_{sp} = \frac{ar_s}{n_q}$, $d_{ps} = \frac{r_s}{n_p}$, $d_{pq} = \frac{1-r_s}{n_{iq}}$。

最后，根据政府公布的每天的确诊人数，选择合适的初始值使得模拟的结果与真实数据尽量拟合，得到疫情之后的预测结果。

模拟预测与参数选取

在模型设计完成之后，利用C-SEIR模型对湖北省确诊人数变化的数据进行拟合，并对未来确诊人数的变化进行预测。

图 8 中包含两条不同的预测曲线（蓝色 / 绿色），其中实线为当天确诊人数，虚线为累计确诊人数。从图中可以看出，尽管两种参数选择在前期都与实际确诊人数（米黄色点）近似曲线吻合，并且拐点时间的预测非常接近，但是最终累计感染人数的预测相差非常大，这表明在疫情上升期进行参数估计得到的结果会有较大的误差。模型根据蓝线进行预测，在第 20 天（2 月 10 日）附近，疫情将出现拐点，到 40 天附近（3 月 1 日），感染人数将达到 5 万，这些都与最终的实际情况基本吻合。

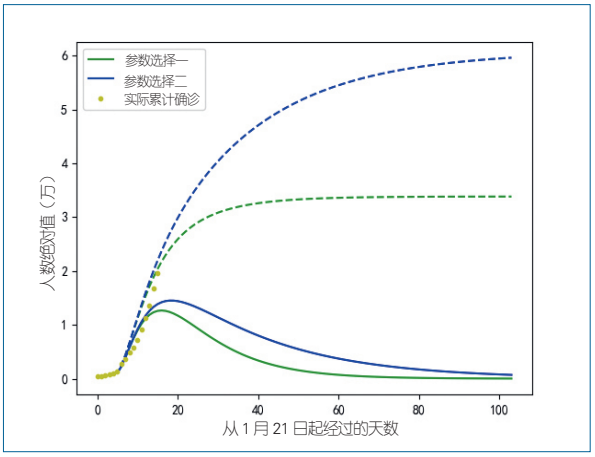


图8 C-SEIR模型预测曲线（基于湖北省确诊数据拟合）

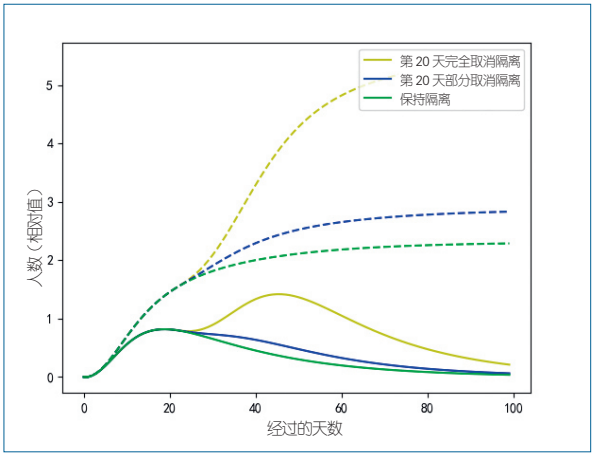


图10 C-SEIR模型预测对比：完全取消隔离会导致新的疫情高峰，而部分取消隔离相比保持隔离则只会导致少量新增人数

模型多样性设计

与基本的SEIR模型相比较，C-SEIR模型所增加的隔离人群更贴近真实情况，而为了更加丰富模型设计的多样性，本文通过不同的方式对现实影响疾病传播的因素进行变量设计。首先是隔离措施的力度，通过隔离者在每日新增确诊患者中被隔离者所占比例来进行描述，隔离者所占比例越高，隔离措施越强。本文也对此变量进行了对比实验，由图9可以看出，新增确诊患者中被隔离者所占比例的高低对疫情峰值时间影响不大，但会显著影响累计患病人数。

模型多样性使得疫情的预测变得更加可控，可以通过改变变量的方式得到不同条件下的预测结果。在实验中，我们采用该模型对解除隔离对疫情感染人数的情况进行了预测分析。如图10所示，如果在单日新增达到峰值（2月10日）时取消隔离措施，疫情会出现第二个峰值（黄线），而部分取消隔离措施（蓝线）相比完全隔离（绿线）则只会导致少量新增感染人数。这样的可调整性也有利于防控措施的展开。

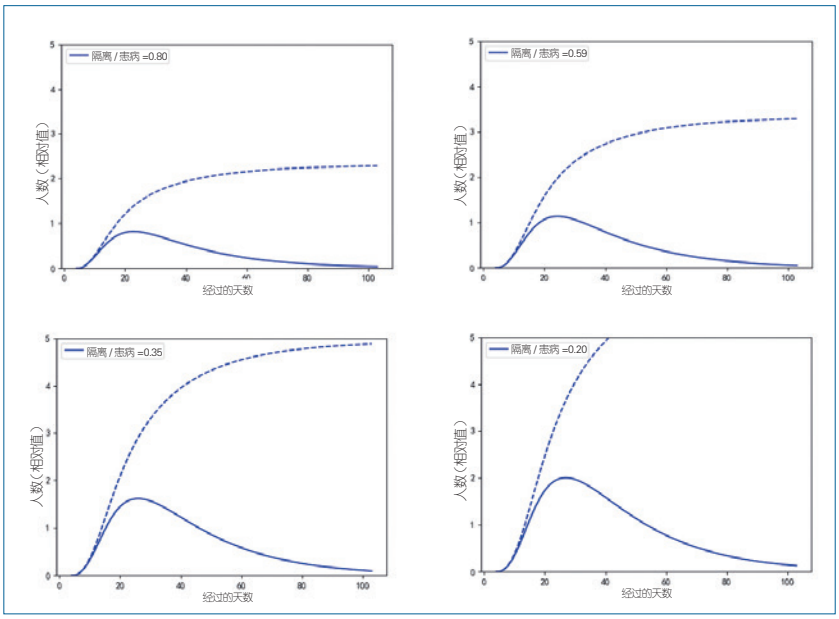


图9 C-SEIR模型预测对比：已隔离人员占患病者的不同比例

针对SEIR模型在分析本次疫情发展时存在的问题，我们使用C-SEIR模型，对人群分类进行进一步细分，并设计了疫情传播对应的状态转移方程，结合疫情传播的分阶段特点和防控措施的影响，我们对模型参数的估计进行了进一步优化。在对疫情发展的模拟中，我们的模型结果与实际数据基本一致。同时通过模型中特殊变量的设计，可以模拟封城、多强度隔离等措施，大大提高了模型的多样性，提高了在预测过

程中的可控性。

总结

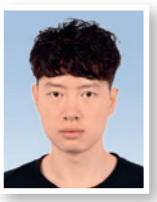
本文针对本次新冠肺炎疫情传播，通过科学可视化分析，发现疫情初期感染情况、疫情中期感染情况和春运期间人口流动数据之间的联系，对疫情发展做出分阶段传播的假设；同时，通过对比各省市感染疫情人数的变化情况，得出本次疫情发展具有区域间差异的结论。

本文针对 SEIR 模型的四类人群分类存在的不合理情况，利用 C-SEIR 模型，增加新的人群分类，并结合可视化发现的疫情传播特点以及隔离等防控措施的影响，对疫情传播进行更合理的建模，模型的实验模拟结果与实际的疫情发展情况基本相符。同时，模型的预测具有多样性，可以通过变量的调整得到不同条件下的预测结果，为疫情防控提供更合理的指导。 ■

致谢：感谢 C-SEIR 的作者张娟教授对我们模型研究提供的讨论反馈。

陈宝权

CCF 会士、理事，*CCCF* 专题主编。北京大学博雅特聘教授、前沿计算研究中心执行主任。北京市未来影像高精尖创新中心首席科学家。IEEE Fellow。主要研究方向为计算机图形学与数据可视化。
baoquan@pku.edu.cn



史明镒

CCF 学生会员。山东大学交叉研究中心硕士研究生。主要研究方向为姿态估计、动画生成、深度学习。
irubbly@gmail.com

蒋鸿达

CCF 学生会员。北京大学前沿计算研究中心博士研究生。主要研究方向为动画生成、相机控制、深度学习。
jianghd@pku.edu.cn

其他作者：倪星宇　阮良旺　姚贺源　王梦迪　宋振华　周　强　葛　彤

参考文献

[1] Wu J T, Leung K, Leung G M, et al. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study[J]. *The Lancet*, 2020, 395(10225): 689-697.

[2] 2019 新型冠状病毒（2019-nCoV）疫情状况的时间序列数据仓库，https://github.com/BlankerL/DXY-2019-nCoV-Data.

[3] 百度地图慧眼 - 百度迁移 [OL]. http://qianxi.baidu.com/.

[4] Zhang J , Lou J , Ma Z , et al. A compartmental model for the analysis of SARS transmission patterns and outbreak control measures in China[J]. *Applied Mathematics and Computation*, 2005, 162(2):909-924.

CCF 会员与分部工委布局会员服务

2020 年 6 月 22 日，CCF 就会员发展和服务策略进行了讨论，CCF 秘书长**杜子德**，CCF 会员与分部工委主任**吴国斌**、主任助理**童咏昕**，CCF 会员部部长**戴丽霞**参会，CCF 会员与分部工委副主任**李贝**进行了书面发言。

会议认为，导致目前会员黏性不高的主要原因是会员参与度不高，而让会员参与是对会员最好的服务方式，也是让会员认同并留在学会的最好方式。要最大限度地让会员动起来，一方面要对会员、分部、学生分会“自编自演”的节目加以鼓励和表彰；另一方面，总部也要设计一些激励会员参与互动的活动。

会议重点围绕如何策划并组织让会员动起来的产品、如何为非 985 和非 211 类高校提供定制化服务等议题进行了深入讨论。基于本次会议的基本思路，CCF 会员与分部工委将组织 CCF 各分部代表进一步讨论。