基于 vSEIdRm 模型的人口迁移以及离汉交通管控对新冠肺炎疫情发展的影响分析 *

顾嘉 陈松蹊 董倩 邱宇谋

摘要：不同于传统(Susceptible-Exposed-Infected-Removed) SEIR 流行病传播动力学模型，本文在近期研究的 Varying Coefficient Susceptible-Exposed-Infected-Diagnosed-Removed（vSEIdR）模型基础上加上人口迁徙 (Migration) 模块，设计开发了 vSEIdRm 模型，该模型考虑跨区域人口迁徙对疫情传播的影响，并允许流行病传播参数随时间变化。本文首先对人口迁移数据进行统计分析，建立其与各省新冠肺炎疫情发展的联系。之后我们基于此模型估计了疫情初期各省来自武汉的输入病例数，并定量刻画了离汉交通管控的效果。研究结果显示，离汉交通管控措施有效地减少了各省的疫情规模。

关键词：流行病动态传播模型；人口迁徙；离汉交通管控影响

The effect of population migration and Wuhan lockdown on the control of COVID-19 epidemic based on vSEIdRm model

Jia Gu Songxi Chen Qian Dong Yumou Qiu

Abstract： Different from traditional (Susceptible-Exposed-Infected-Removed) SEIR epdidemic model, based on the Varying Coefficient Susceptible-Exposed-Infected-Diagnosed-Removed (vSEIdR) model in our previous study, in this paper we add a population migration compartment and porpose the vSEIdRm model, which takes the effect of cross-regional migration on the epidemic into consideration and allows the parameters to vary with time. We first conduct statistical analysis on the population migration data and connects the migration and the progression of the COVID-19 epidemic. Further, based on the new model, we estimate the would-be imported cases to the provinces from Wuhan, Hubei in the absence of Wuhan

* 基金项目: 本研究得到自然科学基金项目“面向管理决策大数据分析的理论与方”（92046021），“变系数流行病学模型的统计推断”（12071013）和“面向儿童脑发育障碍性疾病的神经机制建模与辅助诊疗算法”（12026607）的资助，和北京大学统计科学中心和数量经济与数理金融教育部重点实验室（LMEQF）的帮助。

lockdown which quantifies the effect of the Wuhan lockdown. Our results show that the Wuhan lockdown effectively reduced the size of the epidemic in other provinces。
Keywords: Epidemic dynamic transmission model; Migration; Effect of Wuhan lockdown

一、引言

北京时间 2020 年 1 月 23 日 10 时起，武汉市的公交、地铁、轮渡、长途客运暂停运营；机场、火车站离汉通道关闭 ------ 这意味着为期 76 天的，一次对千万级别大城市采取的严厉封城措施的开始。截止到 1 月 24 日 12 时，湖北共有 13 个城市区域公共交通停运，它们分别是武汉、鄂州、仙桃、枝江、潜江、黄冈、赤壁、荆门、咸宁、黄石（含大冶市、阳新县）、当阳、恩施、孝感。截至 3 月 20 日 24 时，全国累计确诊人数达 81,008 例，其中武汉市 50,005 例，湖北省除武汉 17,795 例，中国大陆余下 30 省（不含港澳台、湖北）累计确诊人数 13,208 例。疫情很大程度上被限制在湖北省乃至武汉市内。

一个在控制病毒传播层面很自然的问题是如何量化评估武汉及湖北封城对各省市疫情发展的影响？ 时值春运高峰，如果当时没有采取封城措施，巨大的人口流动又会对疫情起到什么作用？ 有学者将人口的迁徙与新冠肺炎的时空分布建立联系，提出了一套疫情风险指标模型 （ Jia 等（2020））； Zhou 等 (2020) 考虑了人口流动限制对于新冠肺炎疫情在深圳传播的影响；Zhang 等 (2020) 探究了人口迁移对中国国内疫情传播的作用，并发现了两者之间非常强的正相关性；Cao 等（2020） 分析了导致中国疫情发展的系统性因素，并得出了人口迁徙是主要原因的结论。还有学者探究了疫情在中国发生的最初 50 天内，管控措施起到了什么样的作用 （Tian 等（2020）），他们的结论是离汉交通管控导致疫情到达全国其他城市的时间推迟了 2.91 天。北京大学陈松蹊课题组就新冠肺炎疫情提出了基于变系数 SIR 模型 （Varying Coefficient Susceptible-Infected-Removed model），并且估计了中国大陆各省的传染病再生系数R_t，及时评估疫情发展变化（Sun 等（2020））； 他们之后又提出了 vSEIdR 模型并用于分析比较 25 国疫情控制效果，得出了疫情较晚爆发的国家并没有吸取到较早爆发国家的经验的结论（Gu 等

（2020））。面对侵袭全球的第二波乃至第三波疫情，他们将无症状感染者纳入模型考虑，提出vSIADR 模型，并结合线性混合效应模型分析各国管控措施对于疫情防控的效果 （Yan 等（2021））。

目前为止关于人口迁徙对新冠肺炎传播的相关工作, 比如 Kraemer 等 (2020) 的人口迁徙数据来源是百度迁徙指数 （qianxi.baidu.com），这只是反映人口多少的相对指标。在本文中，我们将使用来自国家统计局（中国联通智慧足迹提供）的 2020 年 1 月 1 日至 2020 年 4 月 10 日以及 2019 年 1 月 20 日至 2019 年 4 月 10 日的人口迁徙数据，其包含从湖北省和武汉市到全国其他 30 个省市自治区（不含港澳台，下同）的每日人口双向流动。我们利用提出的变系数 SEIdRm 模型（vSEIdRm），对传统 SEIR 流行病传播模型做出推广，将武汉市到其他各省的人口迁徙对各省市新冠肺炎疫情发展的影响考虑在内，并基于此模型进行情景分析，量化人口迁徙对各省疫情发展的影响以及离汉交通管控的效果。若 2020 年 1 月 23 日武汉市未采取封城措施，则对于从 1 月 24 日起四周内（至 2 月 20 日）已经达到 250 例确诊病例的 17 个省市，其总计确诊病例将增加至 27,963 例（95%置信区间：25,127 – 31,159 例），确诊总人数增加比例达到 151%（95%置信区间：126% - 180%）。这意味着在一轮疫情开始的初期，及时切断病毒传播路径（采取封城等措施），对有效控制疫情有重要作用，这也为有效防控冬季第二轮疫情反弹提供了经验。

本文的结构安排如下：第二部分是对人口迁移数据的描述性统计分析，探究武汉市到各省的人口迁移以及封城与各省确诊人数的相关性；第三部分建立 vSEIdRm 模型并提出估计方法，同时给出情景模拟结果；第四部分在实际数据上进行情景分析，定量评估离汉交通管控的效果；第五部分是讨论与总结。

二、数据汇总与描述性统计分析

（一）数据

本文使用的人口移动数据集包括了 2020 年 1 月 1 日起至 2020 年 4 月 10 日武汉市和湖北省向其余 30 个省级行政单位的每日人口流动，武汉市向湖北省其余各市的流动，以及全国各省之间

的每日人口流动信息。数据还包括了 2019 年 1 月 20 日至 2019 年 4 月 10 日的相应人口移动数据。按照农历，2020 年 1 月 23 日（离汉交通管控日）是大年二十九，其对应 2019 年 2 月 3 日，而 2019 年 1 月 20 日对应春节前两周，使用自 2019 年春节前两周的数据是用于评估离汉交通管控效果，分析若不封城情况会如何变化。本文使用的每日新增确诊病例、累计确诊、现存病例、治愈及死亡数据来自丁香园新型冠状病毒肺炎疫情实时动态数据 (https://ncov.dxy.cn/ncovh5/view/pneumonia)。

（二）描述性统计分析

通过对离汉交通管控前后两周（以 2020 年 1 月 23 日为分隔点）从湖北以及武汉到全国其余 30 个省（区、市）（不包含香港、澳门和台湾省）平均人口流动和 2019 年农历同期的平均人口流动数据分析可以发现与 2019 农历同期相比，2020 年人口流动显著减少（见附件图一和图二所示）。通过测算离汉交通管控前后从湖北省向全国各省份的人口流动的改变率（分别计算 2020 年 1 月 23 日 24 点前两周和后两周的平均流动，以此计算改变率），可以发现除安徽省外（68%）所有省份都达到了 80%以上的下降率，其中四川、北京、福建、河北、贵州、广西、山西、辽宁、云南、甘肃、海南、吉林、内蒙古、宁夏、黑龙江和青海这 16 个省份下降率超过 90%。结合离汉交通管控前后从湖北省向全国各省份人口流动的改变率测算结果来看，封城措施对限制人口流动确实起到了显著作用（见附件图三）。同时，由于封城措施主要在武汉市实行，并逐步扩展到全省，可以发现如果以湖北省全体来看，封城前两周时间内湖北省往外省流动总计超过 1040 万人，封城后两周内仍然有超过 127 万人流出湖北，下降约 88%；相比较之下封城前两周武汉市往外省流动总计超过 185 万人，封城后两周内武汉市往外省流动总计约 1.3 万人，下降超过 99%。另外，仅 1 月 22 日零点至 1 月 23 日 24 点，武汉市总计流出人口达到 119 万，其中 21 万流向外省，占前者的 17.6%。

（三）相关性分析

通过对封城前（2020 年 1 月 10 日至 2020 年 1 月 23 日）从武汉市到各省的人口流动（对数尺度下）与 2020 年 1 月 23 日起第一周各省的确诊人数以及截至 3 月 15 号的各省总确诊人数的

分析（见附件图四（a）(b）），结果显示人口流动和各省确诊人数之间有较强的正相关性，Pearson 相关系数（下同）分别为 0.72 和 0.76（对应的单边 t-检验 p 值分别为4.3×10^{-6}和4.9×10^{-7}）。进一步，发现了离汉交通管控前不同时段下人口流动总数与各省累计确诊人数的相关性是随时间而变化的，且可以观察得到两个结论：（i）随时间推移，相关性逐渐上升，但是上升过程不单调：从 1 月 27 日至 1 月 30 日出现了一个约 0.05 的下降，之后基本单调上升；（ii）统计封城前不同时间段的人口流量总数对正相关的结论没有显著影响，并且可以发现 1 月 23 日当天从武汉到达全国各省的人数与最终的各省确诊具有最高的正相关性，比考虑封城前更多天的人口流动总数再计算的相关系数至少超出了 0.02（见附件图五）。

三、模型建立

（一） vSEIdRm 模型

为了考虑武汉市到各省份的人口迁移的作用并做定量分析，我们对传统的 SEIR（Susecptible-Exposed-Infected-Removal）（Hethcote，2000）模型做了推广，提出 vSERIdRm 模型。我们将一个非湖北省市的人群在一个时刻t分为五类状态：易感者$S(t)$, 感染但未出现症状或还未确诊者 $E(t)$, 确诊者 $I(t)$ 和移出者 $R(t)$ (包含康复和死亡)。其中 E和I状态均具有传染性，但是只有I状态的人群是可以被观测到的，即E状态的数据是缺失的。由于使用的是日数据，我们令

$$\frac{dS(t)}{dt} = S(t+1) - S(t), \qquad \frac{dE(t)}{dt} = E(t+1) - E(t),$$
$$\frac{dI(t)}{dt} = I(t+1) - I(t), \qquad \frac{dR(t)}{dt} = R(t+1) - R(t).$$

vSEIdRm 的模型设定是，给定t时刻已有的全状态变量信息$\mathcal{F}_t$，增量$(\frac{dS(t)}{dt}, \frac{dE(t)}{dt}, \frac{dI(t)}{dt}, \frac{dR(t)}{dt})$的条件期望满足以下微分方程：

$$\begin{cases} \dfrac{dS(t)}{dt} = (1 - \ p_t^E)A(t) - \left(\beta_t^I \ I(t) + \ \beta_t^E \ E(t)\right)\dfrac{S(t)}{M}\ , \\ \dfrac{dE(t)}{dt} = p_t^E A(t) + \left(\beta_t^I I(t) + \beta_t^E E(t)\right)\dfrac{S(t)}{M} - \alpha_t E(t), \\ \dfrac{dI(t)}{dt} = \ \alpha_t E(t) - \gamma_t I(t), \\ \dfrac{dR(t)}{dt} = \ \gamma_t I(t). \end{cases} \tag{1}$$

其中$A(t)$代表t时刻从武汉到达该省份的人数，对应的p_t^E代表$A(t)$中感染者所占比例；β_t^I和β_t^E分别代表I和E状态群体的感染力；α_t和γ_t分别代表检出率和移除率，M是该省市的人口总数。由于来自武汉市的人口迁移对各省人口总数影响不大，我们固定总人口数为常数M。定义$N(t) = I(t) + R(t)$为t时刻累计确诊人数。t时刻有效再生数R_t可以定义为 (Gu 等 （2020）):

$$R_t = \left(\frac{\beta_t^E}{\alpha_t} + \frac{\beta_t^I}{\gamma_t}\right)\frac{S(t)}{M}. \tag{2}$$

有效传染再生数R_t代表的是一个染病者平均可以感染的易感者的人数，因此R_t是度量一个地区的传染是在扩张（$R_t > 1$）还是在收缩（$R_t < 1$）的关键指标（Nishiura 和 Chowell （2009））。

我们可以进一步对模型式（1）指定的条件期望建模，具体来说，我们可以考虑条件独立 Poisson 分布。令$\Delta S(t) = S(t+1) - S(t) \ - (1 - \ p_t^E)A(t)$，$\Delta E(t) = E(t+1) - E(t), \Delta I(t) = I(t+1) - I(t), \Delta R(t) = R(t+1) - R(t)$。于是有$\Delta E(t) = \ p_t^E A(t) + \Delta S(t) - \Delta N(t)$，$\Delta I(t) = \Delta N(t) - \Delta R(t)$。由于实际情况中 $\Delta S(t)$, $\Delta N(t)$, $\Delta R(t)$分别代表染病，确诊和康复的过程，生成机制相对独立，因此可以假设它们在时刻 t服从条件独立的泊松分布:

$$\begin{cases} -\Delta S(t) \ \sim Poisson\left\{\left(\beta_t^I \ I(t) + \ \beta_t^E \ E(t)\right)\dfrac{S(t)}{M}\right\}, \\ \Delta N(t) \sim Poisson\{\alpha_t \ E(t)\}, \\ \Delta R(t) \sim Poisson\{\gamma_t I(t)\}. \end{cases} \tag{3}$$

（二）参数估计方法

在模型中，由于$E(t)$不可观测，出于参数可识别性 (Identification) 的考虑，我们设定$\beta_t^I = \frac{\beta_t^E}{r}$，其中$r > 1$是一个待定超参数。$r > 1$是由于在此次新冠肺炎疫情中，一旦某个个体被确诊，

他/她将居家隔离或是在医院隔离治疗，这意味着确诊个体的传染力β_t^I会显著小于确诊前的传染力β_t^E。以下在给定(r、α、p_t^E)的前提下，我们给出估计感染率 β_t^E的方法，之后我们将给出选择(r、α、p_t^E) 的方案。在参数估计中，最主要的难点在于 E状态的人群是不可观测的。然而根据我们的模型假定，我们有$\Delta N(t) \approx \alpha_t E(t)$，于是我们可以得到$E(t)$的一个“估计量”：$\hat{E}(t) = \frac{\Delta N(t)}{\alpha_t}$。于是根据 式（1）中第二条方程，我们可以建立以下关系式：

$$\hat{E}(t+1) - \hat{E}(t) = p_t^E A(t) + \widetilde{\beta_t^E}\left(\frac{I(t)}{r} + \hat{E}(t)\right) - \alpha_t \hat{E}(t), \tag{4}$$

其中$\widetilde{\beta_t^E} = \beta_t^E \frac{S(t)}{M}$根据人口中易感染群所占比例做出调整的传染力参数。我们定义：

$$Y_t = \hat{E}(t+1) + (\alpha_t - 1)\hat{E}(t) - p_t^E A(t)\ ,\quad X_t = \frac{I(t)}{r} + \hat{E}(t)\,. \tag{5}$$

在时刻 t，我们可以用 Y_t对 X_t做局部核回归来获得参数估计 $\widehat{\widetilde{\beta_t^E}}$, 即关于$\beta$最小化以下目标函数

$$\sum_{i=1}^{T}(Y - X_i\beta)^2 B\left(\frac{t-i}{h}\right), \tag{6}$$

其中 B 是一种从通常核函数调整过的边界核函数（Boundary Kernel）(Jones，1993)，h是平滑带宽（bandwidth），以下我们统一使用带宽$h = 7$。使用边界核函数的目的是为了消除在 $t = T$ 附近，即接近边界时的估计偏差（bias）。于是我们可以获得以下核估计量（Kernel Estimator）：

$$\widehat{\widetilde{\beta_t^E}} = \frac{\sum_{i=1}^{T} X_i Y_i\, B(\frac{t-i}{h})}{\sum_{i=1}^{T} X_i^2 B(\frac{t-i}{h})}. \tag{7}$$

令$\hat{S}(t) = N - N(t) - \hat{E}(t)$和 $\widehat{\beta_t^E} = \widehat{\widetilde{\beta_t^E}}\ \frac{M}{\hat{S}(t)}$，从而 $\widehat{\beta_t^I} = \frac{\widehat{\beta_t^E}}{r}$。从式（3）的最后一条方程中我们可以发现 $\mathrm{E}\,[R(t)|\mathcal{F}_t] = \gamma_t I(t)$，于是我们可以用$\Delta R(t)$ 对$I(t)$做不带截距项的局部核回归来获得移出率γ_t的估计：

$$\widehat{\gamma_t} = \frac{\sum_{i=1}^{T} I(i)\Delta R(i)\, B(\frac{t-i}{h})}{\sum_{i=1}^{T} I(i)^2 B(\frac{t-i}{h})}. \tag{8}$$

对于比例r的选取，由于估计方法对参数的选取不敏感 (Gu et al., 2020)，在本文以下的计算分析中，统一采用$r=5$。我们定义有效传染再生系数R_t的估计量$\hat{R}_t$为

$$\hat{R}_t=\left(\frac{1}{\alpha_t}+\frac{1}{r\hat{\gamma}_t}\right)\widehat{\beta_t^E}. \tag{9}$$

（三）参数自助法（Parametric Bootstrap）方法构造参数置信区间

为了对上述参数估计进行统计推断，我们在条件 Poisson 增量下的 vSEIdRm 模型式（3）下，给定利用原始数据得出的“估计量” $(\hat{S}(t),\hat{E}(t),\hat{\beta}_t^E,\hat{\gamma}_t)$，使用自助法构造估计的置信区间。在每一时刻$t$, 依据如下条件 Poisson 模型产生$t+1$天的增量

$$\begin{cases} -\Delta^* S(t)\ \sim Poisson\left\{\left(\hat{\beta}_t^I\ I(t)+\ \hat{\beta}_t^E\ \hat{E}(t)\right)\dfrac{\hat{S}(t)}{M}\right\}, \\ \Delta^*\ N(t)\sim Poisson\{\alpha_t\ \hat{E}(t)\}, \\ \Delta^*\ R(t)\sim Poisson\{\hat{\gamma}_t I(t)\}. \end{cases} \tag{10}$$

根据式（10）的条件 Poisson 分布设定以及给定的初始值 $(\mathrm{S}(0),\mathrm{E}(0),\mathrm{I}(0),\mathrm{R}(0))$，我们可以反复生成传染病传播过程曲线$(S^*(t),E^*(t),I^*(t),R^*(t))$。重复 *B(B=500)* 次上述抽样，将再抽样得到的数据记为$D^b=\{S^{*b}(t),E^{*b}(t),I^{*b}(t),R^{*b}(t)\}_{t=1}^T, b=1,\dots,B$。对每一组数据，我们可以使用 3.2 节给出的参数估计方法得到$(\hat{\beta}_t^{I,b},\hat{\beta}_t^{E,b},\hat{\gamma}_t^b)$。在$t$时刻对再抽样得到的估计关于利用原始数据得到的估计$(\hat{\beta}_t^I,\hat{\beta}_t^E,\hat{\gamma}_t)$做中心化（Centering），使得再抽样估计的均值等于$(\hat{\beta}_t^I,\hat{\beta}_t^E,\hat{\gamma}_t)$。我们定义 t 时刻各参数的 95%置信区间端点分别为为中心化再抽样估计的 2.5%和 97.5%分位数。

（四）模拟试验

如果我们没有把从武汉到各省的人口迁移项$p_t^E A(t)$考虑进来，情况会如何？即模型的错误设定（Mis-specification）会对参数估计造成怎样的影响？在此我们进行两组模拟试验。为了简单起见，所有的参数都被设定为常数, 并按照式（3）所述的方式生成数据，其中

① $M=5\times 10^8$ 代表总人口；$p_t^E\ A(t)=20,\ 50,100$ 代表每日输入病例数。

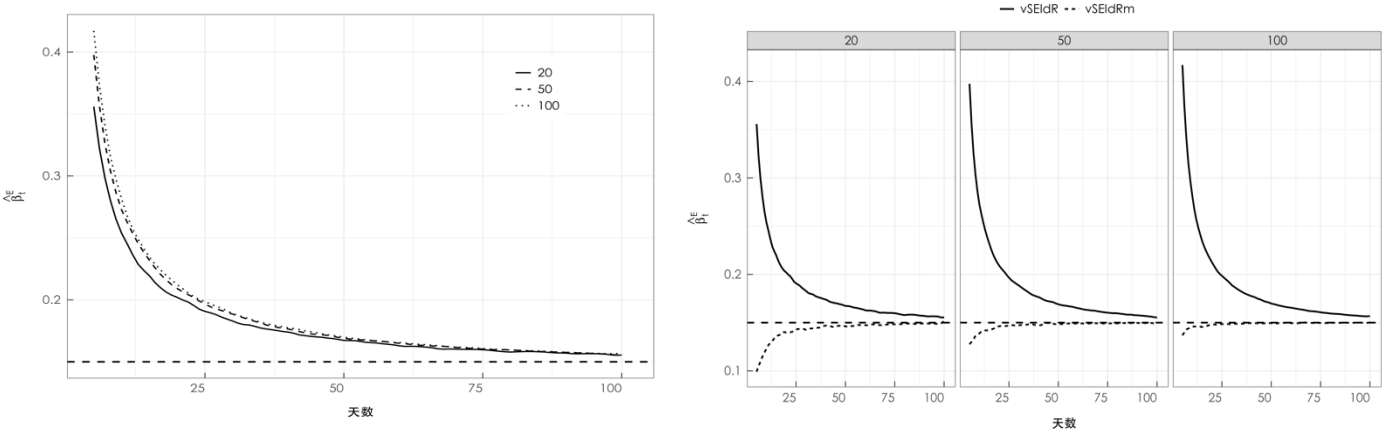
② $E(0)=5, I(0)=5$ 代表初始感染与确诊人数。

③ $\beta_t^E = 0.15$，$\alpha = 0.25$，$\gamma = \frac{1}{21}$ 分别代表感染率（E 状态时），以及确诊率和移出率；$r = 5$，这意味着$\beta_t^E = 5\,\beta_t^I$；在此设定下基本再生数R_0（即 $\frac{S(t)}{M} = 1$时）为 1.23。

对于每个给定的$p_t^E\,A(t)$，我们可以在带条件 Poisson 增量的 vSEIdRm 模型式 （3）下生成数据，再估计参数β_t^E。 重复实验$B = 500$次，将估计的均值曲线绘制成图一。从 图一（a）中我们可以发现，忽略人口迁移 （即在回归中忽略 Y_t中的$p_t^E A(t)$项）会导致早期对参数β_t^E的显著高估，在模拟情形下疫情爆发一周后仍然有超过 100%的高估。随着疫情的发展，当总体确诊人数不断增加，相比之下输入病例作用不太重要时，估计又逐渐收敛到真值 0.15，但是收敛过程非常缓慢，与真值差距缩小到 10%以内需要大约 2 个月（57 天）的时间。

图一（b）展示的是另一组对比，其中 vSEIdRm 和 vSEIdR 分别代表考虑人口迁移和不考虑人口迁移的参数估计方法。可以发现，虽然两种方法都逐渐向真值收敛，vSEIdRm 关于β_t^E的估计方法仅需大约两周时间就能达到 10%以内误差。而此时忽略人口迁移的 vSEIdR 模型仍然有超过 50%的高估。若换算作R_t的值，此时 vSEIdRm 方法的估计值为 1.13，而 vSEIdR 方法估计值为 1.85，与真值（1.23）差异较大。准确掌握真实的有效传染再生数R_t对实际防控疫情时及时掌握疫情剧烈程度有重要的意义。

（a）vSEIdR　　（b）vSEIdRm 和 vSEIdR 比较



图一：忽略人口迁移影响（使用 vSEIdR 模型）在不同日输入病例情形下关于β_t^E的参数估计（a）以及 vSEIdRm（虚线）和 vSEIdR（实线）的比较（b）。虚线代表模拟时的参数真值 0.15。

（五）基于观测数据的 p_t^E估计

由于早期检测效率等原因，虽然我们获取到了从武汉到各省的人口流动的数据$A(t)$，但是其中新冠肺炎病例所占比例却是未知的，因此我们需要基于模型对该参数做出估计。为此我们做出以下假定。令 2020 年 1 月 10 日为 $t=1$，1 月 10 日至 1 月 23 日是离汉交通管控前两周，考虑到病毒的潜伏期及 14 天隔离期 (Lauer 等 （2020）)，所以我们以此两周时间段作为病毒大范围扩散到各省的时间窗口。

1.对某一选定省份 P，以及给定的参数η，令p_t^E服从以下变化模式：

$$p_t^E(\eta)=\begin{cases}\frac{\eta}{100}e^{\frac{R_0}{14}t}, & 1\le t\le 14;\\ 0, & 15\le t.\end{cases} \qquad (11)$$

$p_t^E(\eta)$ 的设定理由如下：t = 14 对应 1 月 23 日，当天上午十时起，离汉交通管控正式实施，因此我们可以近似认为 1 月 24 日之后流出人群都经过了严格的检查，所以我们令这之后的参数$p_t^E=0$。在 1 月 10 日至 1 月 23 日这两周时间内，是病例输出的主要时间段。由于初期对疫情认识不足，在这一时间段内，我们认为武汉经历了一段病毒快速传播的过程，检测率α和移出率γ 都可以近似认为是 0。由于武汉是本次第一轮疫情的中心，没有外来输入病例的影响，因此对于武汉市内这一时间段的新冠肺炎传播，从模型式（1）的第二条方程来看，我们有$\frac{S(t)}{M}\approx 1$，近似地

$$\frac{d\,E(t)}{d\,t}\approx\ \beta_t^E E(t)\frac{S(t)}{M}\approx\ \beta_t^E E(t),$$

因此我们可以得出结论，疫情初期 E 状态人群在武汉市内经历了一段指数增长的过程，这导致武汉市初期向各省的病例输出率p_t^E的指数增长。求解以上微分方程，我们有

$$E(t)=E(0)e^{\beta_t^E t}=E(0)e^{\frac{R_0}{14}t},$$

其中$R_0=\beta^E D$是武汉市此次疫情的基本再生数，我们设定为 5.7（Sanche 等 （2020）），D = 14 是从感染到确诊的平均时间长度。由此我们可以得出前述封城前两周内p_t^E的参数化形式 （11） 。

2. 诊断率α的选取

令α服从以下变化模式：

$$\alpha_t(p^E) = \begin{cases} 0, & 1 \le t \le 10; \\ \dfrac{\dfrac{1}{3.5} - \xi_1(p^E)}{14}(t-11) + \xi_1(p^E), & 11 \le t < 25; \\ \dfrac{1}{3.5}, & t > 25. \end{cases} \tag{12}$$

大多数省份的病例从 1 月 21 号开始出现一个比较明显的增长，并且随着时间推移，检测能力有一个快速上升的过程，因此我们将α_t设定为一个分段线性的函数，并且以 1 月 21 日作为非零α_t起始点。根据钟南山团队发表的文章（Guan 等 （2020）），我们将平稳期的检出率α给定在 1/3.5（即大约 3.5 日从染病到被确诊，这反映了随着疫情发展，检测能力的快速上升）。而关于左端点，即ξ_1，我们可以如下考虑：对某一省份 P，定义 $E_0 = \sum_{t=0}^{13} p_t^E\ A^P(t)$，其中$A^P(t)$代表第 t 天武汉向该省份的人口流动, E_0 可以代表离汉交通管控前向该省输入的病例总数。根据模型式（1）的第 2，3 式，近似地我们有

$$\alpha_{13} \approx \frac{N(14) - N(13)}{E_0}.$$

可以假定封城前 1 月 21 日和 1 月 23 日这两天的检出率差别不大，于是又有$\xi_1 = \alpha_{11} \approx \alpha_{13}$。由此我们将检出率$\alpha_t$定义成为如式（12）所给定的$p_t^E$的函数。

3. 基于确诊曲线$N(t)$拟合效果的参数 $(p_t^E(\eta), \beta_t^E, \beta_t^I, \gamma_t)$ 估计

给定$p_t^E(\eta)$，我们可以首先估计出检出率α_t，然后进一步利用 3.2 小节的估计方法，我们可以得出其余参数估计。于是从 $t = 14$ 起，以 $\big(S(13), E(13), I(13), R(13)\big) = (M\ -\ E_0, E_0 - N(13), N(13), 0)$ 作为初始值，在省份 P 我们可以根据式（1）所确定的增量模型生成一条基于模型的确诊曲线，记为 $N^H(t)$，其中M代表该省份人口总数。 考虑 1 月 24 日 （$t = 15$）起四周的拟合效果，定义如下距离函数：

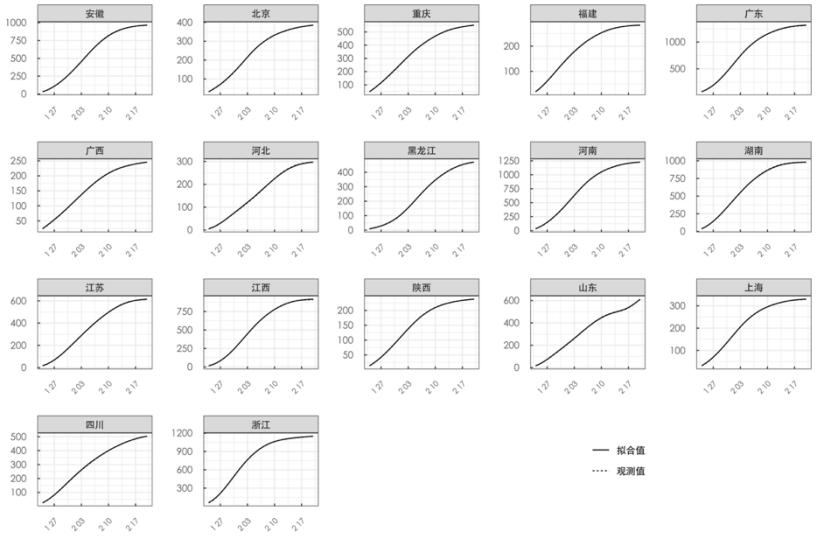
$$D\ (\eta) = \sqrt{\frac{1}{28}\sum_{t=15}^{42}\big(N(t) - N^H(t)\big)^2} \tag{13}$$

在参数值范围$\eta \ge 0$内极小化$D(\eta)$，可以得到相应的参数估计。

四. 疫情初期输出病例估计及情景分析

（一）输入病例估计

我们选取了 17 个截至 2020 年 2 月 20 日确诊病例达到 250 人的省市（分别是安徽、北京、重庆、福建、广东、广西、河北、黑龙江、河南、湖南、江苏、江西、陕西、山东、上海、四川和浙江），根据前文基于距离的估计方法，对相应参数进行了估计，其中确诊曲线的拟合效果展示在图二和表一，参数估计展示在表一。



图二：2020 年 1 月 24 日至 2 月 20 日 17 省确诊人数及相应拟合值的曲线。（该 17 省截至 2020 年 2 月 20 日确诊人数达到 250 人及以上）

首先从图二来看，该方法的拟合效果比较理想，这为我们的估计方法提供了有力地支撑。表一首先给出了各省η取最优值时目标函数式（13）的值，可以发现最优函数值的取值范围是$[0.25,2.21]$，平均值是 1.16（标准误 SE=0.16，下同），这意味着平均意义下拟合值与实际确诊数字每天的误差不到 2 例。与湖北省相邻的 6 个省、直辖市（安徽、重庆、湖南、河南、陕西和江西）用星号标出。其中E_0代表封城前武汉市向各省输出的病例数的估计，N是截至 3 月 15 日各省市总的确诊人数。

表一　　17 省市的拟合效果$D(\eta)$以及相应的参数估计值

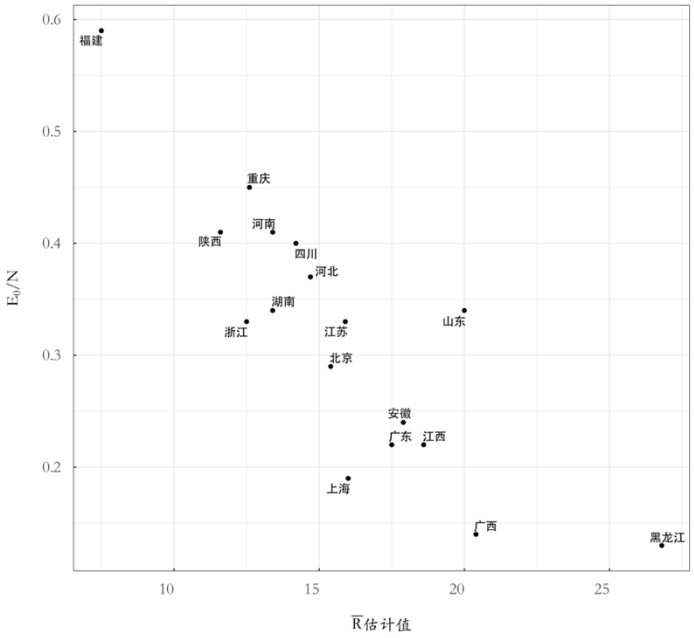
	省份	$D(\eta)$	η	E_0	N	E_0/N比例	$\overline{R_1}$	$\overline{R_2}$
1	福建	0.43	0.0075	169	287	0.59	7.8	7.5
2	陕西*	0.25	0.0149	98	240	0.41	12.3	11.6
3	浙江	1.55	0.0630	395	1193	0.33	12.9	12.5
4	重庆*	1.08	0.0067	253	563	0.45	12.8	12.6
5	河南*	1.91	0.0015	511	1233	0.41	13.9	13.4
6	湖南*	2.17	0.0020	332	986	0.34	14.2	13.4
7	四川	1.06	0.0292	208	523	0.40	14.8	14.2
8	河北	0.95	0.0053	115	308	0.37	15.5	14.7
9	北京	0.59	0.0064	129	448	0.29	15.9	15.4
10	江苏	1.20	0.0054	203	614	0.33	16.9	15.9
11	上海	0.41	0.0138	68	352	0.19	16.4	16.0
12	广东	2.21	0.0061	295	1339	0.22	18.1	17.5
13	安徽*	1.50	0.0025	228	969	0.24	18.6	17.9
14	江西*	1.69	0.0027	205	914	0.22	19.3	18.6
15	山东	1.72	0.0259	251	738	0.34	20.4	20.0
16	广西	0.46	0.0027	35	251	0.14	20.7	20.4
17	黑龙江	0.61	0.0298	60	477	0.13	27.4	26.8
总数				3550	11435			
均值		1.16	0.0075	209	673	0.32	16.3	15.8
(标准误)		(0.16)	(0.002)	(31)	(89)	(0.03)	(1.1)	(1.0)

注：E_0代表封城前武汉市向各省输出的病例数，N 代表截至 3 月 15 日各省确诊人数，比例计算的是两者比值，即输入型病例占各省总体病例的比值。$\overline{R_1}$和$\overline{R_2}$分别代表在 vSEIdR 和 vSEIdRm 模型下计算的累计传染力指标式（14）。与湖北相邻省份用*号标示。17 省市排名按照累积传染力指标$\overline{R_2}$由小到大排列。

如果我们考察比例（E_0/N），这一项越小，代表一个输入病例带来的二次乃至三次传播越严重，换言之也就是疫情控制效果较差。这一点可以通过和有效传染再生数R_t的估计值建立联系而得到验证。令 1 月 10 日为$t=1$，考虑离汉交通管控后四周内（1 月 24 日至 2 月 20 日）R_t曲线下面积，我们如下定义累积传染力指标$\bar{R}$:

$$\bar{R}=\frac{1}{2}\left(R_{15}+R_{28}+2\sum_{t=16}^{27}R_t\right) \tag{14}$$

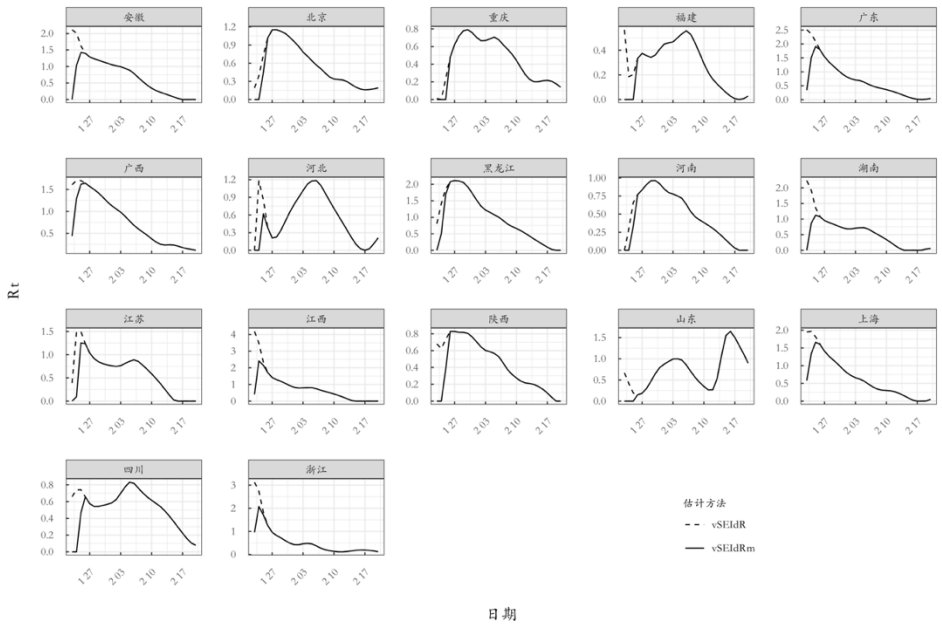
图三显示输入病例比例与累积传染力指标的 Pearson 相关系数达到了-0.85（对应单边 t-检验的 p 值为8.05×10^{-6}）。根据有效传染再生数的定义（2），我们可以认为比例E_0/N是一个直观反映疫情控制效果的度量。



图三：17 省累积传染力指标$\bar{R}$（在 vSEIdRm 模型下计算）与对应输入病例占总确诊数（截至 2020 年 3 月 15 日）比例的散点图。Pearson 相关系数为-0.85，对应的单边 t-检验的 p 值为8.05×10^{-6}。

我们将 17 省分为两组，分别是：(i)与湖北相邻 6 省；（ii）其余 11 省，以考察两组间比例的差异。相邻组的平均比例是 34.5%（SE：4%），其余省份组的平均比例是 30.3%（SE：4%），这意味着对于相邻省份来说，即便面临着更大的输入压力，各省份的管控仍然做的很不错，甚至好于其余省份，不过这一差异在统计意义下并不显著，两样本单侧 t-检验的 p 值是 0.23。所有省份中，福建省的估计比例值最高，达到 58.9%，这意味着该省在这 17 省的疫情防控中，表现最为出色。而在相邻省份组中，我们可以看到江西和安徽省的比例显著的低于其他相邻省（江西、安徽分别为 22.4%和 23.5%，相邻省份除去江西、安徽平均比例 40.2%（SE：2%）），说明这两个省的表现有所不足。这一点也可以从图四的传染再生系数估计曲线中看出：江西的基本再生数R_0估计达到了 2.4；而安徽和湖南虽然R_0值非常接近 (分别是 1.4 和 1.1)，但湖南R_t降低到 1 以下仅仅用了 5 天（1 月 27 日），相比之下，安徽则花了 12 天（2 月 3 日）。17 个省份中共有 5 个省市有效再生数R_t始终没有超过

1，分别为重庆（R_t最高值为 0.79，下同）、福建（0.56）、河南（0.96）、陕西（0.83）和四川（0.83），这意味着疫情在这 5 个省市始终处于有效地管控之下。



图四: 17 省 1 月 23 日起至 2 月 20 日传染病再生系数R_t在两种估计方法（vSEIdR 和 vSEIdRm）下的估计

从估计上来看，17 省中表现最差的是黑龙江省，接近 90%的病例来自本地传播，而这一估计也可以从黑龙江省疾病预防控制中心发布的消息（http://www.chinanews.com/sh/2020/02-07/9082717.shtml）得到侧面验证。

从总量上来看, 从武汉往 17 省的输出病例总数估计值是 3555 例, 占 17 省截至 3 月 15 日总确诊病例数（11435 例）的 32%，这说明近三分之一的病例是输入型病例，这也从侧面验证了离汉交通管控的重要意义，在当代快捷交通运输的背景之下，若没有及时果断的旅行限制，病毒的传播严重程度将不可设想。

我们也可以通过累积传染力指标（14）来对各省的疫情防控效果做出评估，表一分别展示了在 vSEIdR 和 vSEIdRm 模型下计算的累积传染力指标$\overline{R_1}$ 和$\overline{R_2}$。我们发现没有考虑人口移动的累计传染再生数$\overline{R_1}$比考虑了的$\overline{R_2}$平均小 0.56，主要是未将输入病例从总确诊数中扣除所带来的估计偏差。在

vSEIdRm 模型下，累积传染力指标的最小值为 7.5（福建），最大值为 26.8（黑龙江），后者是前者的 3.6 倍。根据这一指标，我们可以对 17 省市控制疫情的表现做出排名：福建、陕西、浙江、重庆、河南、湖南、四川、河北、北京、江苏、上海、广东、安徽、江西、山东、广西和黑龙江。

（二）情景分析

在 2020 年 1 月 23 日离汉交通管控之后，各省市也立刻采取了管控措施。为了能够更好地评价封城的效果，一个自然的问题是：如果当时没有封城，疫情发展会如何变化？情况会有多严重？我们试图通过情景试验模拟来回答这个问题。

1.模拟试验设计

对于一个给定的省市，为了评价离汉交通管控对该省市自封城起四周（2020 年 1 月 24 日至 2 月 20 日）的疫情发展影响，我们设计如下模拟情景：

（1）令向该省从武汉流入的流动人群中的患病比例保持在 1 月 23 日的估计值（即封城前的最高值）水平上，保持两周，再用两周时间逐渐下降到 0，以数学表达式为：

$$p_t^E = \begin{cases} \dfrac{\hat{\eta}}{100} e^{R_0}, & 14 \le t \le 27; \\ \dfrac{\hat{\eta}}{100} e^{R_0}\left(1 - \dfrac{t-27}{14}\right), & 28 \le t \le 41; \\ 0, & t > 41. \end{cases} \tag{15}$$

以上仍以 1 月 10 日为第一天$t = 1$。

(2) 从 2020 年 1 月 24 日起的四周（至 2020 年 2 月 20 日），使用 2019 年农历同期（2019 年 2 月 4 日至 2019 年 3 月 3 日）的人口流动数据替换 2020 年的人口流动数据；

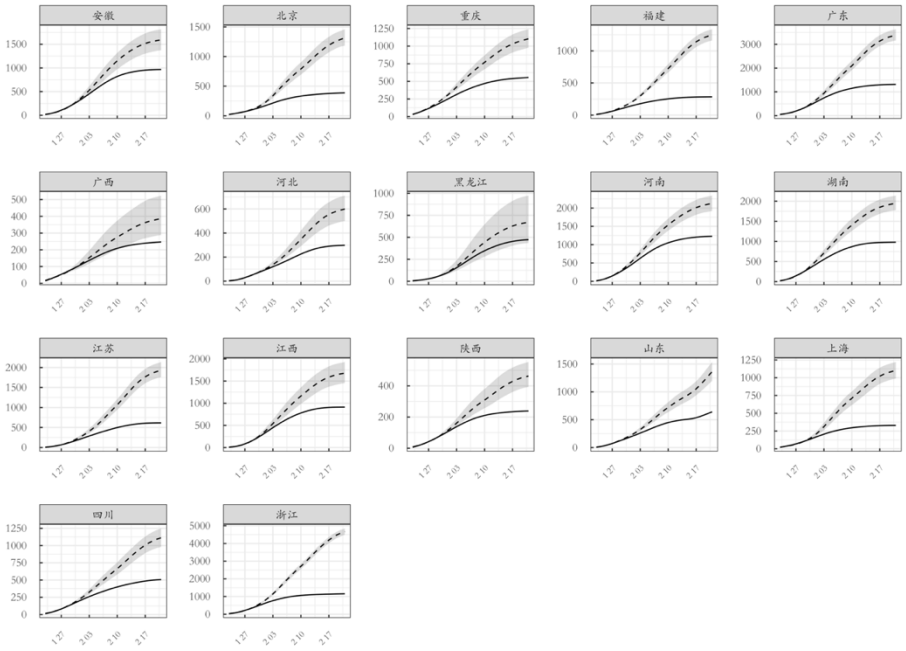
(3) 由于我们主要分析的是离汉交通管控效果的影响，检出率、传染率以及移出率估计值$\hat{\alpha}$，$\hat{\beta}_t$，$\hat{\gamma}_t$均使用该省已经估计的值并保持不变。

2. 结果分析

截至 2020 年 2 月 20 日(离汉交通管控后四周)，17 省市确诊总人数为 11121 人，死亡 81 人。如果在前述的不封城情形下（见图五，表二），确诊总人数将增加至 27963 人（95%置信区间：

25127 – 31159 人），增加 151%，其中来自武汉输入病例为 12508 例，占总确诊人数 43%；死亡人数将增加至 155 人，增加 98%。各省确诊人数增加比例平均值为 154%(标准误：22%)，死亡人数增加比例平均值 162%（标准误：61%；由于实际情形下江苏零死亡，故计算死亡增加比例均值及相应标准误时不考虑在内，下同）。从数值上可以直接得出结论：如果当时没有果断的采取封城隔离的措施，疫情将比后来实际发生的情况严重的多。分省来看，如果当时没有采取封城，北京、福建、浙江、上海、江苏和广东等东部经济大省以及重要城市将受到最为严重的影响, 确诊人数的平均增长比例将达到 259%（标准误：25%），死亡人数平均增加比例为 294%（标准误：80%）。与湖北相邻的 6 个省份管控措施总体表现良好，因此武汉不封城带来的影响相比之下会小一些，但平均来看确诊人数增幅仍然达到了 94%（标准误：10%），死亡人数增幅达到 119%（标准误：26%）。而黑龙江省受到影响相对较小主要是人流量绝对值较小的因素：2019 年 2 月 3 日至 2019 年 3 月 2 日（对应武汉 2020 年 1 月 23 日封城

起一个月）从武汉到黑龙江省的日均人流量是 76 人。但即使是这较小的绝对值，在武汉不封城的情形下，依旧会给黑龙江带来超过 50%的确诊人数增幅（死亡人数增加 26%）。



图五: 17 省确诊曲线观测值（实线）和假想武汉不采取封城时的确诊曲线（虚线）以及相应的 95%置信区间。图中的时间段是 2020 年 1 月 23 日至 2020 年 2 月 20 日。

表二　　17 省截至 2020 年 2 月 20 日实际确诊人数（即封城情形）和武汉不封城情形的确诊人数比较

	省份	确诊人数（封城）	确诊人数（不封城）	增加人数	增加比例（%）
1	福建	283	1259	975	343
2	陕西*	238	485	246	103
3	浙江	1154	4842	3688	319
4	重庆*	554	1109	555	100
5	河南*	1227	2165	937	76
6	湖南*	980	2103	1122	114
7	四川	507	1156	649	128
8	河北	298	615	317	106
9	北京	387	1334	947	245
10	江苏	613	1988	1374	224
11	上海	329	1151	821	249
12	广东	1314	3626	2311	176
13	安徽*	966	1748	781	81
14	江西*	912	1842	930	102
15	山东	641	1394	753	117
16	广西	246	432	185	75
17	黑龙江	472	714	242	51
总数		11121	27963	16842	151

注：与湖北相邻的 6 个省份用*号标出。17 省市排名按照累积传染力指标$\overline{R_2}$由小到大排列。

五、结论

本文分析了人口流动大数据，建立了人口流动与疫情严重程度的相关性，为以后应对此类公共卫生事件提供了参考。进一步，本文拓展了传统的流行病传播模型——SEIR 模型，将确诊前感染和人口迁移带来的跨区域疫情传播纳入考虑，并允许参数随时间变化，构建了 vSEIdRm 模型，使得微分动力学方程对疫情发展的建模更加符合实际情况。本文依据此模型分析了离汉交通管控的效果，得出了如果不封城将造成巨大损失的结论。同时，由于现代交通技术的发达，产生传染病疫情时，受到最严重影响已不一定再是相邻省份，对于经贸联系紧密、人口流动频繁的经济发达省市，必须及早做好响应，执行严格的防疫政策，以免遭受更大的损失。对于经贸联系并不那么紧密或是人口流动相对较少的省市，也不能掉以轻心，正如 Ruan 等(2020) 指出，国家之间（在本次疫情初期即省际）的疫情扩散仅需要少数（5-10）个感染者入境，事实上黑龙江省本轮疫情就经历了一段本地聚集性传播的过程。一次成功的疫情防控需要外防输入和内防本地传播的有机结合。

鸣谢：作者感谢中国联通智慧足迹数据科技有限公司提供的人口迁移数据。

参考文献：

[1] Cao ZC, Tang F, Chen C et al. Impact of Systematic Factors on the Outbreak Outcomes of the Novel COVID-19 Disease in China: Factor Analysis Study[J]. Journal of Medical Internet Research, 2020(22).

[2] Gu J, Yan H, Huang YX, Zhu YR, Sun HX, Qiu YM and Chen SX(2020). Comparing Containment Measures among Nations by Epidemiological Effects of COVID-19[J]. National Science Review, 2020, 7(12), 1847-1851.

[3] Gu J, Yan H, Huang YX, Zhu YR, Sun HX, Xinyu Zhang, Wang YQ, Qiu YM and Chen SX. Better Strategies for Containing COVID-19 Epidemics --- A Study of 25 Countries via an Extended Varying Coefficient SEIR Model. Medrxiv, 2020.

[4]Guan WJ, Ni ZY, Zhong NS et al. Clinical Characteristics of Coronavirus Disease 2019 in China[J]. The New England Journal of Medicine, 2020(382):1708-1720.

[5] Yan H, Zhu YR, Gu J, Huang YX, Sun HX, Zhang XY, Wang YQ, Qiu YM and Chen SX. Better strategies for containing COVID-19 pandemic: a study of 25 countries via a vSIADR model [J]. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2021(477).

[6] Hethcote HW . The mathematics of infectious diseases [J]. SIAM review, 2000, 42, 599–653. doi:10.1137/S0036144500371907.

[7] Jones M. Simple boundary correction for kernel density estimation[J].Statistics and Computing, 1993,3,135–146. doi:10.1007/BF00147776.

[8]Nishiura H and Chowell G. Mathematical and Statistical Estimation Approaches in Epidemiology[M]. 2009. doi: 10.1007/978-90-481-2313-1_5.

[9] Jia J, Xin L, Yun Y , Ge Xu , Jia J, Nicholas C. Population flow drives spatio-temporal distribution of COVID-19 in China[J]. Nature, 2020(582): 1-11.

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

[11] Lauer SA, Grantz KH, Bi QF et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application [J]. Annals of Internal Medicine, 2020, 172(9): 577-582.

[12]Ruan YS, Luo ZD, Wu CI et al. On the founder effect in COVID-19 outbreaks – How many infected travelers may have started them all?[J] National Science Review, 2020.

[13] Sun HX, Qiu YM, Yan H, Huang YX, Zhu YR, Gu J and Chen SX, Tracking Reproductivity of COVID-19 Epidemic in China with Varying Coefficient SIR Model (with discussion)[J], Journal of Data Science, 2020, 18 (3), 455–472.

[14] Sanche S, Lin YT, Xu CG et al. High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2[J]. Emerging Infectious Diseases, 2020, 26(7):1470-1477.

[15] Tian HY, Liu YH, Li YD et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China[J]. Science, 2020(368): 638-642.

[16] Zhang C, Chen C, Shen W et al. Impact of population movement on the spread of 2019-nCoV in China[J]. Emerging Microbes & Infections, 2020, 9(1).

[17] Zhou Y, Xu RZ, Hu DS et al. Effects of human mobility restrictions on the spread of COVID-19 in Shenzhen, China: a modelling study using mobile phone data[J]. The Lancet, 2020.

作者简介：

顾嘉，男，北京大学数学科学学院统计科学中心在读博士，目前为博士二年级。研究方向流行病数据建模与分布式统计推断。

陈松蹊（通讯作者），男，北京大学讲席教授，北京大学光华管理学院、数学科学学院教授。曾是 Iowa State University 统计学终身教授。研究方向为数理统计，超高维统计推断，大数据算法，大气环境统计，计量经济。邮箱：songxichen@pku.edu.cn

董倩，女，统计学博士，现就职于国家统计局统计科学研究所，研究方向为贝叶斯分析及预测、机器学习、数据挖掘、价格指数编制及数字经济测算。

邱宇谋，男，统计学博士，Iowa State University 助理教授，研究方向为大维协方差矩阵统计推断，高维统计推断及其在基因组学中的应用，脑成像统计分析以及因果推断。