

封一

答卷编号（参赛学校填写）：

答卷编号（竞赛组委会填写）：

论文题目：（标明 A、B、C、D... 之一）

A

组 别：（填写研究生、本科生、专科生或中学生）

本科生

参赛队员信息（必填）：

	姓 名	学 号	联系电话
参赛队员 1	王炬皓	2018023935	13170827523
参赛队员 2	姚征兵	2018024041	17773525136
参赛队员 3	孙博	2018024271	18545998714

指导教师：

赵艳芹

参赛学校：

黑龙江科技大学

封二

答卷编号（竞赛组委会填写）：

评阅情况（联赛评阅专家填写）：

联赛评阅 1.

联赛评阅 2.

联赛评阅 3.

新型冠状病毒疫情

摘要

一、问题重述

数据是反映各国疫情防控效果的最真实的指标,对大规模数据的处理、分析、归纳并得出结论是数学模型构建的主要目标,也是各国做出防疫决策的重要手段。各国要做出合理的防控决策关键是要对疾病的社会传播机制构建有效的数学模型从而做出相对准确的预测,并结合各国的国情分析该数学模型对本国的效用和往期阶段决策的滞后影响做出联合评估。

新冠肺炎(COVID-2019)是2019年爆发的全球流行性疾病,对各国社会的经济、秩序和人民的生命财产造成了重大影响,考验着各国对重大公共医疗卫生突发事件的应急能力。COVID-2019的传播在中国已得到显著的控制,但在一些发达国家特别是美国,形势仍旧严重。因为各国的国情差异错综复杂,单一的防控方式

三、模型假设

- 假设 1 当前尚未结束疫情国家的实时数据作为最终数据对模型没有影响
假设 2 在被检测出存在患病情况的患者都会被要求强制隔离,并退出感染系统
假设 3 不考虑各国人种差异造成的新冠感染率差异
假设 4 每个真实染病患者都会去做核酸检测,病症潜伏者只来源于核酸漏检
假设 5 假设在疫情爆发期间,每个易感者的社交范围为 200 到 300 人
假设 6 在新冠肺炎病毒从感染源经历多次传播链,它的传染率也基本保持不变

四、符号说明

4.1 符号说明

符号	符号含义
$S(x)$	以人口为变量的 Sigmoid 函数
S_i	第 i 类国家聚集成团的簇
δ	新冠肺炎的潜伏周期中位数
$I(t)$	t 时刻累计感染人数占人口总数的百分比
$P(y t)$	t 时刻的核酸检测率
$P(n t)$	t 时刻的核酸漏检率
$E(t)$	t 时刻潜伏者人口占总人口的百分比
$population$	泛化指标,在不同的公式中指代相应国家的人口总数
ω_i	与第 i 个国家地理位置和经济发展水平相近的国家构成的集合
ρ_{ij}	i 国和 j 国单日新增感染数据之间的斯皮尔曼相关系数
$s(t)$	t 时刻易感人群占总人群的百分比

$i(t)$	t 时刻已感染人群占总人群的百分比
$r(t)$	t 时刻免疫态人群占总人群的比例
β	新冠肺炎传染率
γ	新冠肺炎感染者进入免疫态的概率
ℓ	新冠肺炎免疫者丧失免疫态的概率
λ	新冠肺炎有效传染率

五、问题的模型建立与分析

5.1 问题一的求解

5.1.1 数据指标的确定

问题一要求自主从网络上采集数据,用来分类世界范围内主要国家的疫情发展特点及抗击疫情状况。截至目前,各个国家的感染情况大致如图 5.1 显示。图 5-1 中各个国家根据感染病例大小按颜色从深蓝到淡黄色渐进变化,白色区域为数据缺失。

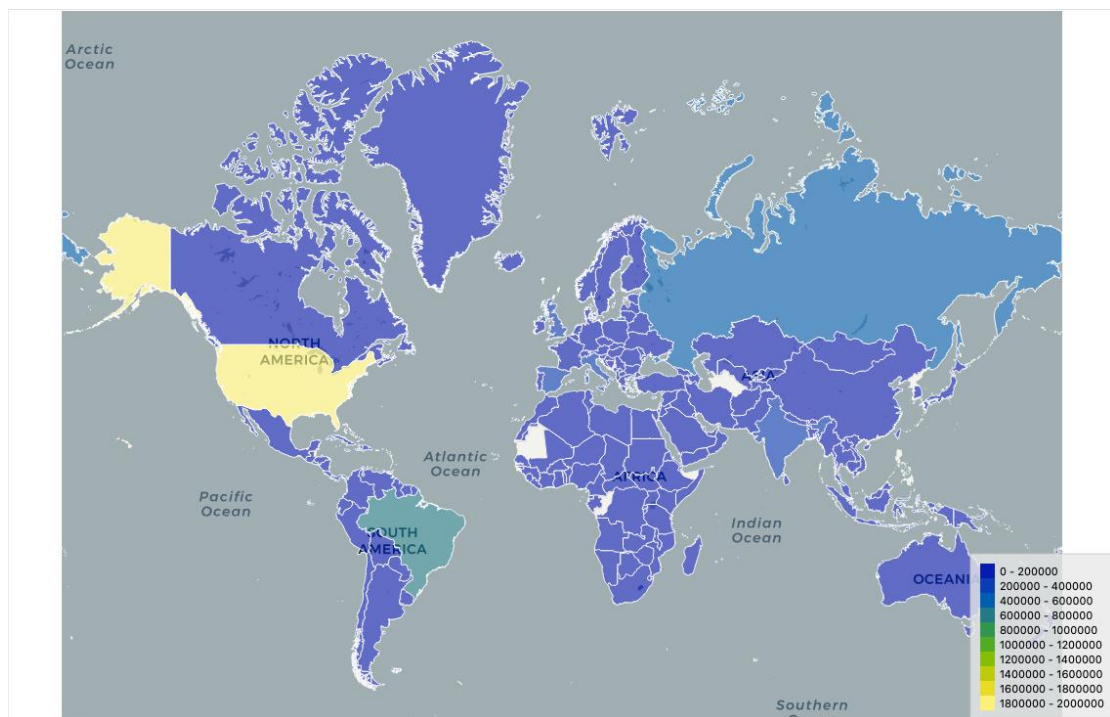


图 5-1 世界感染地图

由于各个国家的政府部门权值结构和基本国情不同,很难从单一数据源来评判疫情的发展情况和抗击疫情的效果。对此收集到了以下几个方面数据从多个维度来反映上述两者的情况:

- (1) 各国最早发现病例时间
- (2) 感染时间跨度
- (3) 最终治愈率
- (4) 最终死亡率
- (5) 单日增长人数峰值占总人口的百分比
- (6) 人口基数
- (7) 医疗水平, 采用 HAQ 指标 (Healthcare Access and Quality Index)

这些数据值量级可能差异巨大, 要对各个维度的数据做 Z-Score 标准化使得原始的数据变为均值为 0, 方差为 1 的分布区间以便做综合考虑。为了简化评估过程, 将数据源 (2)–(7) 作为一个单独的维度——有效政策评估分数的内变量, 这样就可以将各国简化分为四大类:

1. 早发现感染病例, 早制定有效政策
2. 早发现感染病例, 晚制定有效政策
3. 晚发现感染病例, 早制定有效政策
4. 晚发现感染病例, 晚制定有效政策

这四大类可以很容易地根据数据挖掘领域的聚类算法来求得。

5.1.2 数据处理与模型建立

为分析各个国家的有效政策分数, 首先要定义分数的组成:

$$score = effect^p \cdot difficulty^{1-p} \quad (1)$$

$score$ 是政策分数, $effect$ 是政策产生的效果, $difficulty$ 是政策的执行难度, p 是一个大于 0 小于 1 的数。

首先分析政策的执行难度, 一个基本的思想是考虑该国人口越多, 政策越难以执行。在分析政策的执行难度时必须抛弃各个国家的政治制度差异而单单从人口体量来衡量, 这是因为各个国家的政治制度无法量化采集到的数据层面上。而世界上一共有 197 个国家和 36 个地区, 而全球共有接近 80 亿人口, 按照权值国家和地区比为 10:1 来进行计算, 每个国家的平均人口为

$$\frac{80}{197 + 36 \times \frac{1}{10}} \approx 0.657 \text{ 亿人}$$

假设政策的执行难度满足生物学中常用的 S 型曲线, 采用机器学习领域频繁使用的 sigmoid 函数作为拟合曲线, 它的取值为 (0, 1)。原因是 Sigmoid 函数可以作为分类问题的激活函数从而提取隐藏的特征, 而抗疫政策的执行难度是在 5.1 小节中数据 (6) 和 (7) 的一个隐藏特征。Sigmoid 的函数的优点是平滑且易于求导, 它的定义如下:

$$S(x) = \frac{1}{1 + e^{-x}}$$

$S(x)$ 的导数为:

$$\frac{e^{-x}}{(1+e^{-x})^2}=S(x)(1-S(x))$$

取每个国家的人口平均为政策难度执行增长最快的点, 并且人数为 0 时政策的执行难度为零即 $S(0.657)=\frac{1}{2}, S(0)=0$ 改写 sigmoid 函数得到

$$S(x)=\frac{1}{2-e^{-0.657}+e^{-x}}-\frac{1}{3-e^{-0.657}}$$

它在 $(0,+\infty)$ 区间内的取值范围为 $(0,0.272)$ 即政策的执行难度的取值区间, 可对其进行归一化。

再给政策效果 *effect* 给出评分, 由 4.1 小节中的 2), 3), 4), 5), 7) 来衡量。为简化模型可以看成线性关系

$$effect=-\beta_1\bullet t_{span}+\beta_2\bullet c_{end}^*-\beta_3\bullet d_{end}^*-\beta_4\bullet r+C \tag{2}$$

其中 $\beta_1, \beta_2, \beta_3, \beta_4$ 是依据原始数据给出的合理估计值 t_{span} 是各国存在感染病例的时间跨度, c_{end}^* 是最终治愈率和该国医疗水平的比值, 因为该国的医疗水平很大程度上影响了该国的感染治愈率, 而与政策无关, 增加了比值项能消除该国最终治愈率受该国医疗水平的影响。 d_{end}^* 是该国的死亡率与该国医疗水平的乘积, 增加乘积项的原因同上。 r 是各国单日检测病例增长人数峰值除以该国总人口(单位为亿人), 可以很大程度上反映一个国家对于疫情防控措施的及时性和有效性。公式(1)中的加减表示参数与政策效果呈现正相关或是负相关, 此外还加入了一个常数项来限制 *effect* 为一个正数。

根据附件所给的”世界主要国家的疫情发生数据.csv”表格数据, 利用 K-均值聚类算法分类后的国家如表 5-1 所示。

类型	国家
早发现感染病例, 早制定有效政策	China, Ecuador, Equatorial Guinea, Japan, Korea South, Thailand
早发现感染病例, 晚制定有效政策	Albania, Azerbaijan, Belgium, Bosnia and Herzegovina, Canada, Cuba, Denmark, Djibouti, Finland, France, Hungary, Indonesia, Iraq, Ireland, Kosovo, Kyrgyzstan, Moldova, Morocco, Nicaragua, North Macedonia, Panama, Philippines, Poland, Romania, Serbia, Slovakia, Sweden, Turkey, US, United Kingdom, Uzbekistan, Venezuela
晚发现感染病例, 早制定有效政策	Australia, Austria, Costa Rica, Croatia, Czechia, Estonia, Germany, Greece, Iceland, Iran, Israel, Italy, Latvia, Lebanon, Lithuania, Luxembourg, Luxembourg, Netherlands, New Zealand, Norway, Portugal, Slovenia, Spain, Switzerland, Tunisia

晚发现感染病例, 晚制定有效政策	Afghanistan, Algeria, Argentina, Armenia, Bahrain, Bangladesh, Belarus, Bolivia, Brazil, Cameroon, Central African Republic, Chile, Colombia, Congo (Kinshasa), Dominican Republic, Egypt, El Salvador, Ethiopia, Gabon, Ghana, Guatemala, Guinea, Guinea-Bissau, Haiti, Honduras, India, Kazakhstan, Kenya, Kuwait, Maldives, Mali, Mexico, Nepal, Nigeria, Oman, Pakistan, Paraguay, Peru, Qatar, Russia, Saudi Arabia, Senegal, Singapore, Somalia, South Africa, Sri Lanka, Sudan, Tajikistan, Ukraine, United Arab Emirates, Zambia
---------------------	--

表 5-1 经 K-Means 算法分类后的国家分类表

5.1.3 K-Means 模型的解释

K-Means 的算法思想很简单, 仅仅是根据点与点之间的距离大小将给定的数据点划分到 k 个聚类簇中, 以此来达到分类的目的。虽然这个问题在计算上是 NP 困难的, 使用效率较高的启发式算法可以使得结果快速收敛于局部最优解。

K-Means 的目标是找到满足下式满足的聚类 S_i 来满足以下关系

$$\arg \min_S \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2$$

在问题一建立的模型下 $S = (t_{case}^i, score^i)$ 为数据样本在二维平面上的投影, 而是 μ_i 为簇 S_i 集中所有点的均值。

对于已选择 $k = 4$ 的随机生成 4 个初始均值点通过迭代的方式来求得最后的均 μ_{end} 值点和簇团 S_{end}^i 的范围。

K-Means 每次迭代会优化类别的输出簇, 算法按照如下步骤交替进行:

- 分配: 每个均值点周边的观测点分配到聚类中, 使得他们的组内平方和最小。距离的度量采用欧式距离

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \leq \|x_p - m_j^{(t)}\|^2 \forall j, 1 \leq j \leq k\}$$

其中每个 x_p 都只被分配到一个确定的聚类 S^t 中, 理论上它可以被分为更多的类, 但是根据 (4.1) 小节的讨论四大类严格没有重复区间。

- 更新: 对于上一步得到的每一个聚类, 以聚类中观测的图心作为新的均值点

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

算法在观测分配不再变化时收敛, 由于交替进行的两个步骤都会减小目标函数 (1) 的值, 并且分配方案只有有限种, 一定会收敛到局部最优解。

5.2 问题二的求解

5.2.1.1 检测能力评价资源调度

在问题一中, 虽然对各个国家的最终政策得分做出评价, 但是模型基于一个“不太合理”的假设——各国对疫情的抵抗策略相同, 只考虑疫情结束最终结果来对模型做出静态评价, 而尚未考虑疫情发生的动态性和延续性的特点。

为此,在问题二中,采用一个可以泛化的评价指标,即对新冠肺炎的检测能力,来考量一个国家内部应对重大公共医疗卫生突发事件的资源调度能力,如加快投产核酸检测试剂,或新建公共医院等等。

考虑到已感染者区分为已被核酸检测确诊患病和尚未确诊但已患病两部分:

$$I(t) = I(t)P(y|t) + I(t)P(y|t) \quad (3)$$

其中 $P(y|t)$ 是 t 时刻患病人数的检测率,这部分人会被强制隔离,不会进入感染系统, $P(n|t)$ 是未被检测出的患病概率并且 $P(y|t) + P(n|t) = 1$ 在任意时刻成立。

对于采集到的每日感染新增数据,它基于已确诊病例而忽略了漏检病例,因

此可以得到实际的新增病例为 $dI(t) = \frac{\Delta I_{data}^t}{P(y|t)}$, 其中 ΔI_{data}^t 是政府部门在 t 时刻公

布的新增确诊病例。并且对累计潜伏者 $E(t)$ 可以用一个自然潜伏期内新增感染人数乘以漏检测率的积分来确定。因为超过一个潜伏期 δ , 潜伏者或者已经发病住院, 或者已经靠自身免疫力获得治愈, 从而退出感染系统, 可以列出表达式

$$E(t_0 + \delta) = \int_{t_0}^{t_0 + \delta} \beta \left(\frac{1}{P(y|t-1)} - 1 \right) \bullet \Delta I_{data}^t (Population - I(t)) dt \quad (4)$$

其中 $\left(\frac{1}{P(y|t-1)} - 1 \right) \bullet \Delta I_{data}^{t-1}$ 是 $t-1$ 时刻单日新增核酸漏检的潜伏

者, $(Population - I(t))$ 是总易感人群。

在大规模流行病阶段, 疾病的感染趋势是新增感染人数正比于潜伏者人数乘以易感染人数, 潜伏者人数由式子 (4) 的定义。因此根据上述对新增感染人口的假设可以列出如下方程:

$$\frac{\Delta I_{data}^{t+\delta+1}}{P(y|t+\delta+1)} = \beta E(t+\delta)(Population - I(t+\delta)) \quad (5)$$

其中 $E(t+\delta)$ 和 $I(t+\delta)$ 组成了一个动力系统, 对于这个动力系统并不用求解, 只需要查看 $P(y|t)$ 在时间序列上的变化情况。由于 $P(y|t)$ 难以先验确定, 模型采用的一个策略是根据时间的增长将时间序列潜伏周期 δ 为一个窗口拆分成若干个集合并且假设时间窗口内的 $P(y|t)$ 是一个固定值, 拆分过程可表示为:

将时间序列集合 $(0, 1, \dots, t)$ 拆分为 $(0, 1, \dots, \delta+1), (1, 2, \dots, \delta+2), \dots, (t-\delta-1, \dots, t)$

共 $t-\delta$ 个子集, 相应的可以得到 $t-\delta$ 个不同时刻下的 $P(y|t)$ 的值, 如图 5-2。

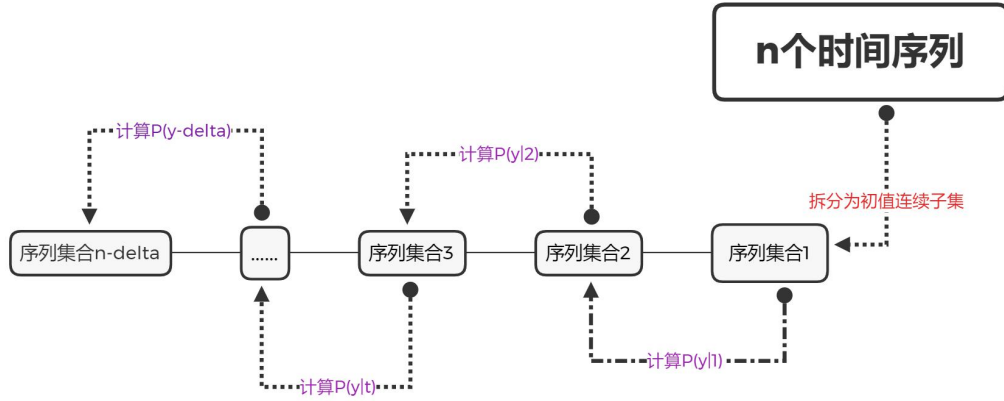


图 5-2 计算 $P(y|t)$ 的流程的简化流程图

根据上述策略, 假定 $P(y|t_0 + \delta + 1) = P(y|t_0 + \delta) = \dots = P(y|t_0) = m$, 并且在 t_0 为时间窗口的初始时刻, δ 为一个潜伏周期内

$$Population - I(t) = Population - \sum_{t=0}^{t_0} \frac{\Delta I_{data}^t}{P(y|t)} - \sum_{t=t_0}^{t_0+\delta} \frac{\Delta I_{data}^t}{m} \quad (6)$$

其中 $\sum_{t=0}^{t_0} \frac{\Delta I_{data}^t}{P(y|t)}$ 这一项中的 $P(y|t)$ 可以用往期计算求得的序列值代入, 用差分方程递归求解。

综合 (5) 和 (6) 不难看出模型对 $P(y|t)$ 概率的确定具有一个潜伏周期 δ 的时滞性。根据编程得到的世界主要国家单日检测率变化如图 5-3 所示。

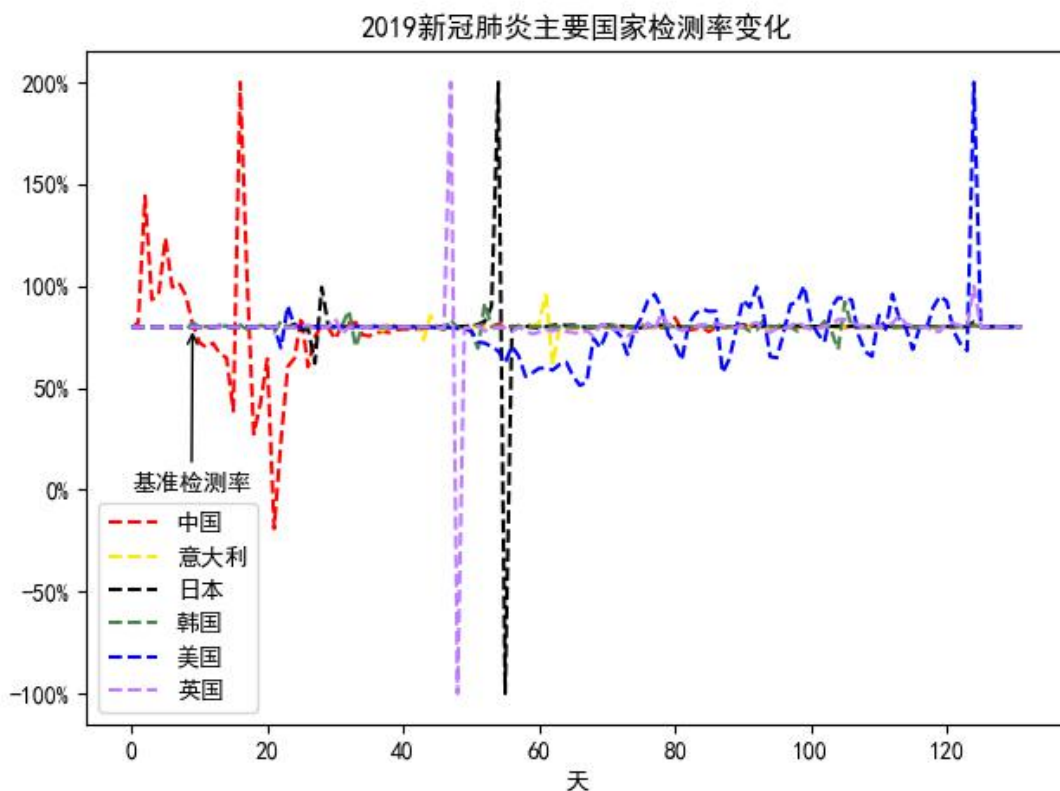


图 5-3 2020 年 1 月 23 日至 6 月 1 日主要国家检测率的变化

在具体的编程环境中采用的接触率感染率 $\beta = 0.01$, 即一个潜伏者在潜伏期内, 平均接触 100 个易感者会致使 1 个易感者感染。并假设在疫情爆发期间, 每个易感者在一个 δ 内的合理社交范围为 200 到 300 人。图中检测率超过 100% 说明历史累积的潜伏者在同一天得到大规模检测, 超过了单日增加的潜伏者人数。发生的原因可能是该国的检测能力呈现大规模增长, 之后检测率回落至 0 以下说明检测的速度还赶不上往期漏检的剩余量。

5.2.1.2 主要六国评价结论

从图 5-3 来看, 从一月份底至三月份初期中国的基础检测率浮动明显, 之后缓慢变动至基准检测率附近。长时间达到基准检测率是满足新增感染人数下降的最低检测率。从图中来看中国检测率爆发式增长的时间为数据记录以来第 18 天左右的时间, 和“世界主要国家的疫情发生数据.csv”表(以下简称数据表)中的 2 月 13 日滞后接近 1 个潜伏期中位数(5.6 天)。意日韩三国从最初的接近无病例到检测率在基准率上下小范围浮动后达到最高值, 清理了历史存留潜伏病例后检测率降至基准线附近, 随后单日新增感染人数逐渐下降。英国的检测率发展趋势和意日韩三国相近, 但英国采取的是“群体免疫”的应对策略, 与意日韩的积极救治不同, 被检测到的潜伏者并不会被强制隔离, 超出了模型的假设范围, 所以导致了显示的检测率与真实的数据差异巨大。美国从数据表记录以来的第 50 天开始检测率变化上下波动较大, 较符合单日新增感染的小峰值情况, 结合数据表中美国感染峰值绝对值大小可以判断其糟糕的抗疫政策。

5.2.2 单日新增病例相关系数评价境外防疫

第 5.2.1 小节中描述了单个国家内部的核酸检测率的变化来评价各个国家面对疫情在公共医疗方面的资源调度情况。本小节基于各国的地理信息和经济发展水平来分析地理距离相近, 经济发展水平相近的国家之间疫情感染数据的相关性检验来反映该国对境外的防控力度, 这是因经济发展水平相近并且地理位置相近的国家往往人口往来频繁, 如中国和日本, 美国和加拿大。因此在没有防控措施的实施下境外输入病例传染本地人口的可能性也就更高。

考虑境外输入病例对潜伏者的影响, 设 ω_i 是第 i 个国家满足上述条件的相似国家构成的集合, 因此可以改写式 (4), 第 i 个国家第 $t_0 + \delta$ 时刻潜伏者人数为:

$$E^i(t_0 + \delta) = \beta \left(\int_{t_0}^{t_0 + \delta} \left(\frac{1}{P(y^i | t-1)} - 1 \right) \Delta I_{data_i}^t + \sum_{j \in \omega_i} k_{ji} \left(\frac{1}{P(y^j | t-1)} - 1 \right) \Delta I_{data_j}^t \right) \quad (7)$$

其中 $\left(\frac{1}{P(y^i | t)} - 1 \right) \Delta I_{data_i}^t$ 是本国 t 时刻存留的潜伏者, $\sum_{j \in \omega_i} k_{ji} \left(\frac{1}{P(y^j | t-1)} - 1 \right) \Delta I_{data_j}^t$ 是从国家集合 ω_i 输入 i 国的潜伏者的求和, k_{ji} 是 j 与 i 国的交通关联系数。结合式子 (7) 和式子 (5) 可以猜测 $\Delta I_{data_i}^t$ 和 $\Delta I_{data_j}^t$ 之间存在相关性, 但并不是简单的线性相关性。

对于非线性相关的两对数据集, 本文采用斯皮尔曼相关系数来量化 $\Delta I_{data_i}^t$ 和 $\Delta I_{data_j}^t$ 的相关性。斯皮尔曼系数具有”非参数”的特性, 这主要反映在两个方面, 首先, 当 X 和 Y 的关系是由任意单调函数描述的, 那么斯皮尔曼相关系数就是完全皮尔逊相关的。其次, 斯皮尔曼相关性系数不需要先验知识便可以准确获取 X 和 Y 的采样概率分布。

斯皮尔曼相关系数被定义为等级变量之间的皮尔逊相关系数, 对于样本容量为 m 的样本, m 个原始数据 X_i, Y_i 被转换等级数据 x_i, y_i , x_i 和 y_i 分别是 X_i 在集合 X 中的排名和 Y_i 在集合 Y 中的排名。斯皮尔曼相关系数 ρ 由被观测的两等级变量之间的差值 $d_i = x_i - y_i$ 决定

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (8)$$

在本题所论述的范围, 具体的做法是先根据经纬度地理信息和各国经济发展水平 EDL (Economic development level), 下文用 EDL 来指代。 EDL 可作为单一的数据来源获取, 结合在地球表面上两国家的经纬度信息来计算两国家的距离,

假设两国家的中心点所在的经纬度分别为 $\phi_i, \gamma_i, \phi_j, \gamma_j$, 对于地球可近似认为是半径为 6378km 的完美球体, 两国家中心点的距离公式为

$$\begin{aligned}\Delta\phi &= \frac{\phi_j - \phi_i}{180\pi} \\ \Delta\gamma &= \frac{\gamma_j - \gamma_i}{180\pi} \\ R &= 6378\text{km} \\ \text{distance} &= 2R \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\gamma}{2}\right) + \sin^2\left(\frac{\Delta\phi}{2}\right) \cdot \cos(\gamma_1) \cdot \cos(\gamma_2)}\right)\end{aligned}\quad (9)$$

j 国对 i 国之间的人员流动相关性可表述成

$$Cor_{ij} = \frac{ELD_i / ELD_j}{dis \tan ce_{ij}} \quad (10)$$

公式(9)表明 j 国对 i 国之间的人员流动相关性并不等于 i 国对 j 国之间的人员流动相关性。

结合上述对 Cor_{ij} 的讨论对于 ω_i 的选择可定义为

$$\begin{aligned}\omega_i &= \{j\} \\ s.t. \\ j &\in \text{Max}_k \{Cor_{ij}\}\end{aligned}$$

ω_i 中集合元素 j 的个数可由给定的参数 k 来任意更改。然后计算 i 国和集合 $j \in \omega_i$ 之间单日新增感染之间的斯皮尔曼相关系数, 即在式(8)前的讨论中定义 $X = \Delta I_{data_i}^t, Y = \Delta I_{data_j}^t$, 并用式(8)进行计算并给出各国境外防疫的评价。

5.3 问题三的求解

5.3.1 复杂网络概述

现代的交通网络让传染病的传播越来越呈现无标度网络的特性, 新冠肺炎在初期的传播网络具有小世界、无标度、高聚类的典型特征, 一个典型的无标度网络如图 5-3 所示。本小节将会基于 SIRS 类传染病的传播过程来对各个国家的历史数据进行建模。

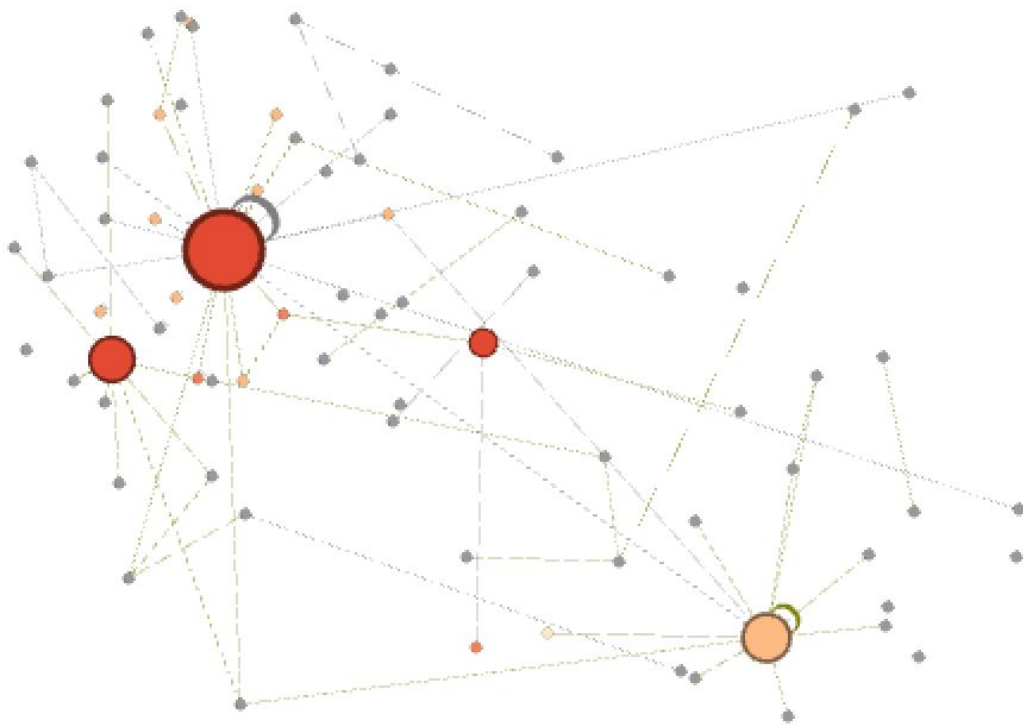


图 5-3 一个简化的表示复杂网络的拓扑图结构

众多的研究表明:对于在均匀网络的传播过程存在一个临界阈值,超出这个临界阈值,传染病就会趋向于大规模传播。SIRS 模型将感染人群分为三类:易感态(S),感染态(I)和免疫态(R)。易感人群如果与感染人群接触会以概率 β 被感染,感染者以概率 γ 治愈并进入免疫态,免疫者则以概率 ℓ 丧失其免疫性。分别用 $s(t), i(t), r(t)$ 来表示不同时间下三类个体的人群占总人群的比例。利用平均场理论来模拟三类个体在复杂网络上传播的动态变化过程能得出整个网络的传播阈值特性。

5.3.2 非均匀网络状态下的 SIRS 传染模型

在复杂网络中,将每个独立的个体视作一个节点,如果个体和个体之间存在接触,那么就用节点之间的连线来表示,称作度。用 $\theta(t)$ 来表示 t 时刻任意一条已存在的边与感染者的连接概率。若用 $\langle k \rangle$ 来表示整个传染网络的平均度,对于节点度相关的复杂网络,任意一条边具有度为 m 的概率与其期望函数 $mP(m)$ 成正比,其中 $P(m)$ 为节点的度分布。因此在一个已知度中, $\theta(t)$ 表示与该节点度不相关的度组成

$$\theta(t) = \frac{\sum_k kP(k)i_k}{\sum_s sP(s)} = \frac{1}{\langle k \rangle} \sum_k kP(k)i_k \quad (11)$$

利用对 $\theta(t)$ 的定义, 可列出节点度为 k 的非均匀网络下 SIRS 的传染方程组

$$\begin{aligned} \frac{ds_k(t)}{dt} &= -\beta k s_k(t) \theta(t) + \ell r_k(t), \\ \frac{di_k(t)}{dt} &= \beta k s_k(t) \theta(t) - \gamma_i(t), \\ \frac{dr_k(t)}{dt} &= \gamma_i(t) - \ell r_k(t). \end{aligned} \quad (12)$$

在稳态条件下 $\frac{ds_k(t)}{dt} = 0, \frac{di_k(t)}{dt} = 0, \frac{dr_k(t)}{dt} = 0$, 从而得到方程组 (12) 中第二个方程组的稳态解为

$$i_k(\infty) = \frac{\beta k \ell \theta(\infty)}{\gamma \ell + (\gamma + \ell) \beta k \theta(\infty)} \quad (13)$$

其中 $\theta(\infty)$ 满足自相容条件

$$\begin{aligned} \theta(\infty) &= \frac{\sum_k kP(k)i_k(\infty)}{\langle k \rangle} = \\ &= \frac{\sum_k k^2 P(k) \frac{\beta \ell \theta(\infty)}{\gamma \ell + (\gamma + \ell) \beta k \theta(\infty)}}{\langle k \rangle} = f(\theta(\infty)) \end{aligned} \quad (14)$$

$\theta(\infty) = 0$ 是式 (14) 的一个平凡解, 另外可从状态迁移方程式 (15) 下得到

$0 < \theta(\infty) \leq 1$ 的稳态解。

$$\theta(t+1) = f(\theta(t)) \quad (15)$$

假设 $\theta(0) = 0 + \varepsilon$, ε 是一个极小量, 因为 $\theta(0) = 0$ 是方程 (14) 的一个平凡解

$$\theta(0) = f(\theta(0)) = f(f(\theta(0))) = \cdots f^n(\theta(0)) = 0$$

用 $f^i(\theta(\infty))|_{\theta(0)=0+\varepsilon}$ 来表示系统在初始条件 $\theta(0) = 0 + \varepsilon$ 下第 i 时刻的 f 函数嵌套层数, 它在数值上就等于 i 。根据状态迁移方程可知

$$\begin{aligned}\frac{d}{d\theta} f(\theta(\infty))|_{\theta(0)=0+\varepsilon} &= \frac{d}{d\theta} f^1(0+\varepsilon) \cdots = \frac{d}{d\theta} f^n(0+\varepsilon) \\ \theta(\infty)|_{\theta(0)=0+\varepsilon} &= 0+\varepsilon + \frac{d}{d\theta} f^n(0+\varepsilon)\end{aligned}$$

定义一个随 t 增长严格递增的函数 $k(t)$, 发现在式 (15) 条件下, 只有当

$$f\left(\frac{d}{d\theta} f(0+\varepsilon)\right) > \frac{d}{d\theta} f(0+\varepsilon) \quad (16)$$

时满足

$$\begin{aligned}f^n\left(f\left(\frac{df(0)}{d\theta}\right)\right) &= f^n\left(\frac{df(0)}{d\theta} + k(0)\varepsilon\right) = f^{n-1}\left(\frac{df(0)}{d\theta} + k(1)\varepsilon\right) \\ &= \cdots = f\left(\frac{df(0)}{d\theta} + k(n-1)\varepsilon\right) = f(\theta(\infty))|_{\theta(0)=0+\varepsilon}\end{aligned}$$

因此在此条件下 $\theta(\infty) = k(n)\varepsilon$, 其物理意义是系统经历 n 个迁移状态后最终达到稳态, 并且根据式 (11) 可限定 $0 < \theta(\infty) \leq 1$ 。将式 (16) 代入到式 (14) 得到非均匀网络状态下的 SIRS 传染模型的状态迁移的条件 $\frac{d}{d\theta} f(\theta(\infty))|_{\theta(\infty)=0} > 1$, 对式 (14) 基于 θ 进行求导整理得到

$$\frac{\beta \langle k^2 \rangle}{\gamma \langle k \rangle} > 1 \quad (17)$$

从而有效传播率 $\lambda = \frac{\beta}{\gamma} > \frac{\langle k \rangle}{\langle k^2 \rangle}$ 为疾病大规模流行的阈值条件。

在式 (17) 中 $\langle k \rangle$, $\langle k^2 \rangle$ 分别是整个传染系统节点之间一次度的期望和二次度的期望, 二次度是指初始节点经过一个媒介顶点连接到终止节点。式 (17) 表明当疾病传染网络的结构很复杂, 节点和节点之间的联系密切, 那么 $\langle k^2 \rangle$ 即节点的二次度的期望会远远大于度的期望, 除非此时治愈率 α 远大于感染率 β , 否则疾病总会爆发为世界性疾病。

查询文献资料显示, 新冠肺炎的各国平均基本可再生数 $R_0 = \lambda = 3.28$, 中位数是 2.79。这意味着要满足新冠肺炎不大规模流行的必要条件是 $\frac{\langle k \rangle}{\langle k^2 \rangle} > 3.28$, 要求着整个复杂网络的图拓扑结构是节点连接稀疏的, 即每个易感者人对传染源的中间媒介感染很稀疏, 最主要的新增感染病例来自于对直接传染源的一次感染。

如果一个网络具有短的平均路径长度和高的聚类系数,则称该网络具有小世界性,随着现代化交通工具的大规模使用,人们的活动半径开始急速增长,现实生活中的“六度分离理论”甚至已经变成“三度分离”,即任意两个素不相识的人通过朋友的朋友,平均最多通过两个人就能彼此认识,以此带来的后果是网络上的行为可以很快的通过这些长程链接扩散到整个网络。因此,现实世界中在各国不进行对感染病例强制隔离、限制交通和人员流动的情况下 $\langle k^2 \rangle$ 往往远远大于 $\langle k \rangle$,这意味着即使 R_0 很小,流行病也有可能大规模传播。

5.3.3 基于 SIRS 模型进行数据建模预测

在 5.3.2 小节,具体讨论了新冠肺炎爆发的阈值条件,给定了新冠肺炎在现实世界中大规模流行的分析结果。根据理论的分析,预测各国疫情的发展趋势可以通过分析该国的历史增长病例来确定 $\langle k \rangle$ 和 $\langle k^2 \rangle$ 并比较 $\frac{\langle k \rangle}{\langle k^2 \rangle}$ 和 $R_0 = 3.28$ 来确定该国的疫情发展趋势是好转还是恶化。

从复杂网络的视角观察 5.2.1 小节的论述,可以将式子(5)改写为

$$\left(\frac{1}{P(y|t)} - 1\right)\Delta I_{data}^t (1 + \beta \langle k(t) \rangle + \beta^2 \langle k^2(t-1) \rangle) = \frac{\Delta I_{data}^{t+1}}{P(y|t+1)} \quad (18)$$

式(18)可以依据 5.3.2 小节已经解出的 $P(y|t)$ 的序列数据来计算 $\langle k \rangle$ 和 $\langle k^2 \rangle$ 在式中舍去了传染链条大于 3 的情况,因为在复杂网络中从初始节点出发,经过三段联系后达到的节点重复性大大增加,且 $\langle k^3 \rangle$ 的系数 β^3 是一个非常小的小数,传染链条大于三的情况基本可以忽略不计。

从复杂网络的理论中可以确定 $\langle k \rangle$ 和 $\langle k^2 \rangle$ 的渐进关系表达式,假设新冠肺炎的二次传染链覆盖的总人数为 N ,在没有终止重复节点的情况下容易得到 $N = [\langle k \rangle + (\langle k \rangle - 1)^2 + 1]$,因此某个人接触某个人的概率为 $\frac{\langle k \rangle}{N}$ 。考虑二次传染链二次传染具有重复节点的数目为 $\langle k \rangle \left(\frac{1}{N - \langle k \rangle} - \frac{1}{N}\right)$,根据节点个数的不变性

$$\langle k \rangle + \langle k^2 \rangle - \langle k \rangle \left(\frac{1}{N - \langle k \rangle} - \frac{1}{N}\right) = N \quad (19)$$

将 $N = [\langle k \rangle^2 - \langle k \rangle + 2]$ 代入式(19),移项整理得到

$$\langle k^2 \rangle \sim \langle k \rangle^2 - 5\langle k \rangle + 14 \quad (20)$$

由于 $[\langle k^2 \rangle] \leq N$,即 $\langle k \rangle^2 - 5\langle k \rangle + 14 \leq [\langle k^2 \rangle - \langle k \rangle + 2]$,因此满足二次传

染链存在重复的内禀条件为最小一次度的期望 $\langle k \rangle \geq 3$ 。

故在现实世界人与人之间接触复杂的情况下, $\langle k \rangle$ 远大于 3, $\langle k^2 \rangle$ 可完全由式

(20) 右侧的式子来替代, 结合新冠肺炎的可再生数 $R_0 = 3.28$, 可根据

$$\frac{\langle k \rangle}{\langle k \rangle^2 - 5\langle k \rangle + 14} = 3.28$$

来计算出阈值 $\langle k \rangle =$

七、参考文献

- [1] 约翰斯•霍普金斯大学新冠肺炎资源中心. 地图&趋势. 世界感染地 [DB/OL].
[2020-06-01]. (2020-06-01)

<https://coronavirus.jhu.edu/map.html>.

[2] Fraser, Christophe, et al. "Factors that make an infectious disease outbreak controllable." *Proceedings of the National Academy of Sciences* [J]101.16 (2004): 6146-6151.

[3] 刘军伟. 地球表面上两点之间距离的几种求算方法[J]. 许昌学院学报, 2007, (12):126-129.

[4] 夏承遗, 刘忠信, 陈增强, 袁著祉. 复杂网络上带有直接免疫的 SIRS 类传染模型的研究[J]. 控制与决策, 2008, (04):468-471.

[5] Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of travel medicine*.