arXiv:2601.04644v1 [stat.AP] 8 Jan 2026

基于聚类的贝叶斯 SIRD 建模：印度水痘流行病学

Nayana Mukherjee[1] 和 Chitradipa Chakraborty[2]

摘要

本研究提出一种基于聚类的贝叶斯 SIRD 模型，利用1990年至2021年的数据分析印度水痘（水痘）的流行病学特征。我们采用年龄结构化方法，将人口划分为青少年、成年人和老年人，以捕捉疾病在不同人口群体中的传播动态。该模型纳入了 Holling 型发病率函数，用于考虑患病率较高时传播的饱和效应，并运用贝叶斯推断估计关键流行病学参数，包括传播率、恢复率和死亡率。研究还进一步开展聚类分析，根据印度各地区水痘传播动态的相似性识别区域聚类，所采用的指标包括发病率、患病率和死亡率。我们运用 K 均值聚类识别出三种不同的流行病学模式，它们在暴发潜力和年龄特异性动态方面各不相同。研究结果表明，青少年是传播的主要驱动群体，而老年人承担着不成比例的高死亡负担。我们的研究结果凸显了针对年龄群体的干预措施的重要性，并表明公共卫生策略应考虑区域异质性，以控制疾病。该模型提供了一个透明、可复现的框架，用于理解长期传播动态，并支持印度水痘防控的循证规划。模拟研究进一步验证了该模型的实际效用。

关键词： 年龄结构化方法；贝叶斯推断；水痘；聚类分析；SIRD模型；传播动力学

[1]数学系， ´马欣德拉大学 Ecole Centrale 工程学院，印度特伦甘纳邦海得拉巴 500043。邮箱：nayana.mukherjee@mahindrauniversity.edu.in[2]北京复杂系统拓扑统计与应用重点实验室，北京数学科学与应用研究院，中国北京 101408。邮箱：chitradipachakraborty@gmail.com（通讯作者）

1 引言

传染病建模已成为现代流行病学中不可或缺的工具，使研究人员能够重建传播动力学、量化流行病学参数并评估长期疾病趋势 [3, 18]。数学模型将生物学假设转化为定量框架，用以描述病原体在宿主群体中的传播 [14, 21]。在众多可用方法中，分舱模型因其清晰性、易解释性和分析上的可处理性而尤为有力。通过将人口划分为离散的健康状态，并利用微分方程组描述这些状态之间的转变，这些模型能够捕捉感染、康复和免疫的基本机制。

由 Kermack 和 McKendrick（1927年）提出的经典易感者–感染者–康复者（SIR）模型 [17] 仍然是数学流行病学中最具影响力的框架之一。在 SIR模型中，个体依次经历三个流行病学状态：易感者（S）、传染者（I）和康复者（R）。各个区室的变化率由一组耦合的非线性常微分方程（ODE） [16],决定。

$$\frac{dS}{dt} = -\beta \frac{SI}{N}, \quad \frac{dI}{dt} = \beta \frac{SI}{N} - \gamma I, \quad \frac{dR}{dt} = \gamma I,$$

其中，β 为传播率，γ 为康复率，$N = S + I + R$ 表示总人口规模。双线性感染发生项 $\beta SI/N$ 表示在均匀混合假设下新感染的发生率，也就是说，每个易感者与传染者接触的概率相同。康复项 γI 表示传染者通过康复并获得免疫或接受隔离，从传播链中移除。尽管 SIR模型十分简单，

它仍能捕捉流行病的基本特征，例如感染波的兴起与消退，以及流行病阈值的存在。它还直观地揭示了感染传播与康复过程之间的平衡。然而，SIR模型假设所有被移除者都会康复，这对于死亡率较高的疾病并不现实。对于这类情况，易感–感染–康复–死亡（SIRD）模型通过明确纳入疾病导致的死亡，对 SIR系统进行了扩展 [5]。在该模型中，传染者可能以 γ 的速率康复，也可能以 μ的速率死于该疾病，从而得到扩展后的系统

$$\frac{dS}{dt} = -\beta\frac{SI}{N}, \quad \frac{dI}{dt} = \beta\frac{SI}{N} - (\gamma+\mu)I, \quad \frac{dR}{dt} = \gamma I, \quad \frac{dD}{dt} = \mu I.$$

因此，SIRD模型在保持质量守恒 $N = S + I + R + D$的同时，区分了康复和死亡结局。参数 β、 γ和 μ 代表决定流行病 [2, 4, 10]发展速度和规模的核心生物过程。对于采用标准质量作用发病率的SIR和SIRD框架，由这些参数推导出的一个关键流行病学量是基本再生数 $R_0 = \frac{\beta}{\gamma+\mu}$。[8, 9, 28], 是衡量传播潜力的简明指标。这个表达式表明， R_0 并非固定输入，而是潜在流行病学速率的结果。它共同取决于新感染发生的速度（通过 β）以及传染者被移除的速度（通过 $\gamma + \mu$）。从概念上说， $R_0 > 1$ 表示流行病可能侵入的条件，而 $R_0 < 1$ 则意味着感染将会消退。实际进行参数估计时，重点在于 β、 γ和 μ，随后根据这些推断出的量计算 R_0 [34]。这些参数的微小变化就可能导致流行病发展轨迹出现显著差异，这凸显了依据数据进行准确估计的必要性。

我们的建模流程从构建SIRD系统开始，该系统采用霍林型饱和发病函数 [15]。我们确保所有参数采用一致的时间基准，且模型变量在量纲上保持齐次。使用与所选时间步长相匹配的常微分方程求解器对系统进行数值积分，并采用后文讨论的鲁棒优化技术估计参数 β、 γ和 μ 。

传染病建模面临的挑战之一，是不同地区、不同人口群体乃至特定年龄组内部的传播模式存在异质性。为应对这一问题，聚类分析已成为识别人口中具有不同流行病学特征的亚组的重要工具。聚类分析假设观测到的人口由若干亚组构成，且不同亚组之间的结局存在显著差异。这种方法尤其适用于对纵向结局进行聚类，即考察个体的轨迹，并依据这些轨迹进行分组。一般而言，纵向结局聚类主要有两种方法：基于模型的方法和基于算法的方法。在基于模型的方法中，会采用有限混合模型，其中每个混合成分对应一个特定的簇。这种方法包括基于组的轨迹模型 [23, 24], ，其中同一簇中的个体遵循相同的轨迹。另一种方法是潜在类别混合效应模型 [22, 27] ，该模型假设个体特异性轨迹围绕共同平均轨迹变化。另一方面，基于算法的方法依靠计算算法进行聚类，而不对数据集的概率分布作出假设。 k均值聚类 [13], 、层次聚类 [36], 和基于相关性的聚类[25]等方法均属于这一类。虽然这些方法广泛用于单变量纵向结局的聚类，但近期研究提出了用于多变量纵

向数据 [6, 7, 19]的贝叶斯潜在类别模型。本研究将聚类分析与SIRD模型结合，旨在识别印度境内水痘传播动态相似的区域聚类。这种方法有助于深入了解疾病负担在年龄特异性和地区层面的差异，为制定有针对性的公共卫生干预措施提供依据。该模型在贝叶斯框架下，使用JAGS中实现的马尔可夫链蒙特卡罗（MCMC）方法进行估计 [26]。

本文的研究动机源于印度有长期水痘（varicella）数据可供利用，涵盖1990—2021年。水痘由水痘-带状疱疹病毒引起，是一种传染性极强的疾病，主要影响儿童，但也可能使成年人和老年人出现更严重的后果，因此适合作为年龄结构化传播分析的案例研究 [35]。近期研究表明，印度境内水痘传播存在显著的区域差异 [31, 33]。年龄特异性因素也会显著影响水痘流行病学，因为与儿童相比，成年人和老年人的死亡率更高 [20]。本文的主要贡献是将聚类与年龄结构化SIRD模型相结合，以刻画印度各地水痘传播动态的区域异质性。不同于以往独立分析各地区或假设传播参数均一的研究，我们联合建模年龄特异性SIRD动态，同时将印度各邦和中央直辖区聚类为具有相似流行病学模式的亚组。这种方法能够以数据为依据识别传播、康复和死亡特征相近的地区，同时考虑时间依赖性和年龄分层。

本文其余部分安排如下。第2节介绍所提出的建模框架，包括年龄结构化SIRD模型的构建，以及用于捕捉传播动力学中区域异质性的聚类方法。

第3节对水痘数据集进行实证分析，并解读研究结果。第4节总结模拟研究的结果。第5节给出一些结论性意见。

2 模型与方法

本研究使用全球疾病负担（GBD）数据集中1990年至2021年的年度水痘数据，构建了一个建模框架。该数据集涵盖印度28个邦和中央直辖区；由于德里的社会人口学和流行病学特征独特，因此将其单独处理。数据集按三个年龄组提供发病率、患病率和死亡数信息：青少年（0–17岁）、成年人（18–59岁）和老年人（$60+$ 岁）。建模时，人口按比例缩放至 $P = 100{,}000$的标准基数。

2.1 采用Holling型发病率的年龄结构化SIRD模型

在该年龄结构化SIRD模型中，每个年龄组的个体被分为四种健康状态：易感者（S）、感染者（I）、康复者（R）和死亡者（D）。易感者与感染者接触后可能受到感染。传播概率取决于不同年龄组之间的互动频率、当前感染者人数，以及一种饱和效应；这种效应确保感染者人数增加时，感染风险不会无限增长。该饱和效应采用Holling型函数形式建模。感染后，个体可能康复，也可能死于该疾病。

为捕捉水痘的流行病学传播动力学，我们采用了

离散时间、年龄结构化 SIRD（易感者–感染者–康复者–死亡者）模型。人口分为三组：青少年（j）、成年人（a）和老年人（o）。每组又进一步划分为四个舱室：

$S_k(t)$（易感者）　I_k（t）（传染者）　R_k（t）（康复者）　D_k（t）（疾病死亡者），

其中 $k \in \{j, a, o\}$。各年龄组的存活人口为 $N_k(t) = S_k(t) + I_k(t) + R_k(t)$。

年龄组之间的传播由霍林型发病率函数控制，该函数考虑了患病率较高时的接触饱和。组 k 所经历的感染力定义为

$$\lambda_k(t) = \sum_{k' \in \{j,a,o\}} \beta_{kk'} \, \frac{I_{k'}(t)/N_{k'}(t)}{1 + \alpha_{kk'} \, I_{k'}(t)/N_{k'}(t)},$$

其中，$\beta_{kk'}$ 表示组 k' 中的传染者向组 k中的易感者传播的强度，$\alpha_{kk'} \geq 0$ 量化了饱和程度。

2.2 控制方程

模　由以下常微分方程组描述 ：

$$\begin{aligned}
\frac{dS_k}{dt} &= -\lambda_k(t) \, S_k(t), \\
\frac{dI_k}{dt} &= \lambda_k(t) \, S_k(t) - (\gamma_k + \mu_k) \, I_k(t), \\
\frac{dR_k}{dt} &= \gamma_k \, I_k(t), \\
\frac{dD_k}{dt} &= \mu_k \, I_k(t),
\end{aligned}$$

对每个 $k \in \{j, a, o\}$，参数 γ_k 和 μ_k 分别表示年龄组 k的康复率和疾病诱发的死亡率。

令 $s_{m,k,t}$, $i_{m,k,t}$, $r_{m,k,t}$ 和 $d_{m,k,t}$ 分别表示 m 州在 k 年龄组于 t年的易感者、感染者、康复者和累计死亡舱室中人口所占的比例。模型的初始值设为 $s_{m,k,1} = 0.97$, $i_{m,k,1} = 0.02$, $r_{m,k,1} = 0.01$, $d_{m,k,1} = 0$.。舱室动力学定义为新增感染$_{m,k,t} = \beta_{m,k} s_{m,k,t} i_{m,k,t}$、新增康复$_{m,k,t} = \gamma_{m,k} i_{m,k,t}$、新增死亡$_{m,k,t} = \mu_{m,k} i_{m,k,t}$，并按以下方式更新

$$
\begin{aligned}
s_{m,k,t+1} &= \max(s_{m,k,t} - \text{new_inf}_{m,k,t}, 0), \\
i_{m,k,t+1} &= \max(i_{m,k,t} + \text{new_inf}_{m,k,t} - \text{new_rec}_{m,k,t} - \text{new_death}_{m,k,t}, 0), \\
r_{m,k,t+1} &= r_{m,k,t} + \text{new_rec}_{m,k,t}, \\
d_{m,k,t+1} &= d_{m,k,t} + \text{new_death}_{m,k,t}.
\end{aligned}
$$

其中，$\beta_{m,k}$ 表示各州和年龄组的传播率，$\gamma_{m,k}$ 表示康复率，$\mu_{m,k}$ 表示死亡率。第 k 个年龄组的基本再生数计算如下：

$$
R_{0,k} = \frac{\beta_k}{\gamma_k + \mu_k}.
$$

2.3 贝叶斯计算

我们采用贝叶斯框架来估计模型参数。年龄和州特异性参数的先验分布设定为 $\beta_k \sim$ Gamma(2,0.2) $\gamma_k \sim$Gamma(2, 1), 和 $\mu_k \sim$ Gamma(1,100). 为处理漏报和测量误差，我们引入了额外的报告因子，其先验分布分别为 $\rho_{\text{inc}} \sim$ Uniform(0,5), $\rho_{\text{prev}} \sim$Uniform(0, 5)), $\rho_{\text{death}} \sim$ Uniform(0.5,2), 以及离散参数 $\sigma_{\text{prev}} \sim$ Uniform(0.001, 0.5)。

SIRD分舱通过以下似然函数与观测数据相关联：

$$y^{\text{inc}}_{m,k,t} \sim \text{Poisson}(\rho_{\text{inc}}\text{new_inf}_{m,k,t}),$$
$$y^{\text{prev}}_{m,k,t} \sim \text{Normal}(\rho_{\text{prev}} i_{m,k,t}, \sigma^2_{\text{prev}}),$$
$$y^{\text{death}}_{m,k,t} \sim \text{Poisson}(\rho_{\text{death}}\text{new_death}_{m,k,t}).$$

对发病率、患病率和死亡率进行联合建模，可提升对传播动态的推断能力。发病率反映新感染率，患病率反映当前感染者所占比例，而死亡数据则提供感染相关死亡率的信息。这种整合既能利用流量数据（发病率），也能利用存量数据（患病率、死亡数），同时考虑测量误差和报告偏倚。我们使用 RJAGS 对模型参数进行估计，并以相应的后验中位数作为估计值；根据样本分位数计算 95% 可信区间。通过轨迹图和计算尺度缩减因子 [11] 来评估链的收敛情况。

2.4 K均值聚类方法

图1展示了随机选取的十个州中，青少年、成年人和老年人的死亡数、发病率及患病率的纵向轨迹。三个年龄组呈现出明显不同的时间模式，而且各组内不同州之间也存在显著差异。这种异质性表明，各州的疾病负担并非以统一的方式演变。数据中呈现的多样化趋势表明，存在具有不同流行病学行为的潜在亚组，因此有必要采用聚类来揭示这些隐藏模式。这种方法有助于更好地理解水痘数据中的年龄特异性和州特异性动态。为捕捉异质性

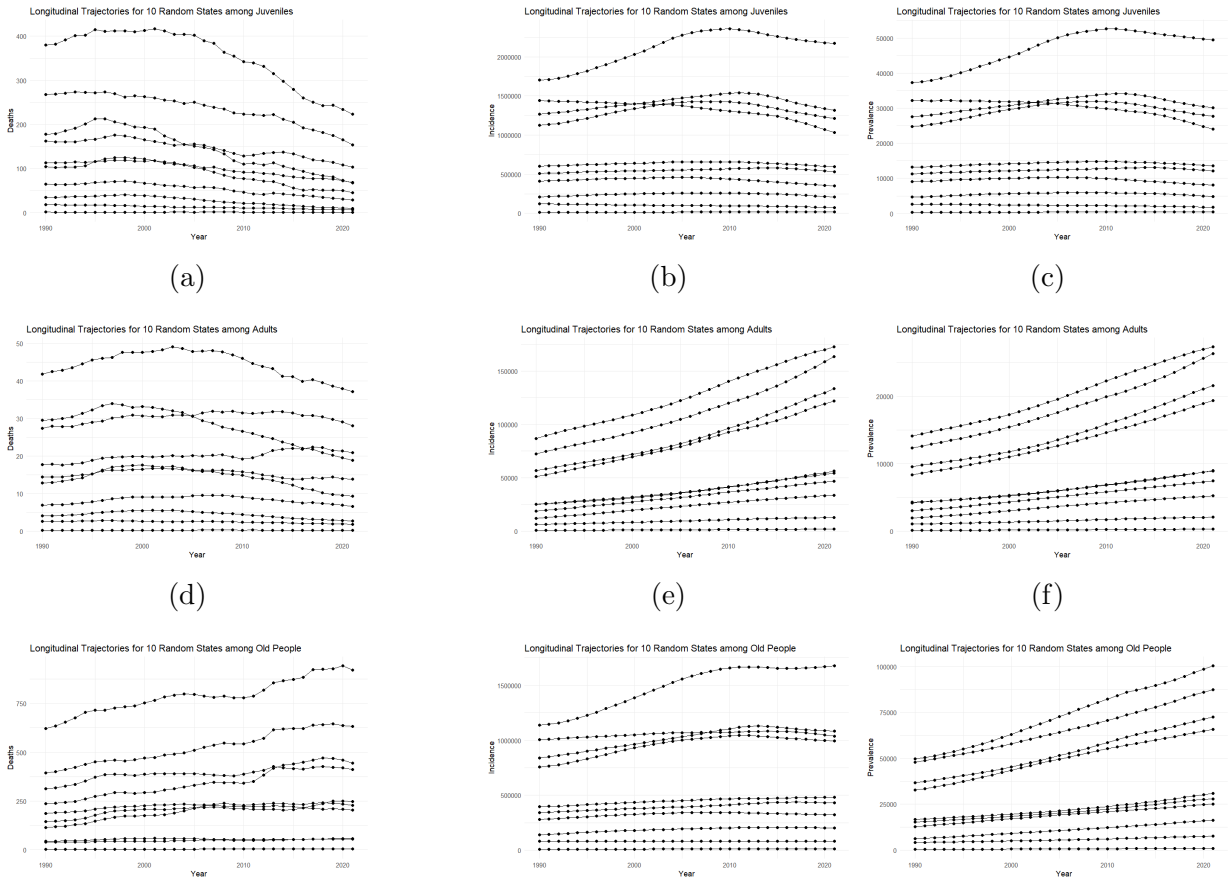


图1：水痘数据集中随机选取的10个州的三项特征的平均纵向轨迹。(a–c) 青少年的死亡数、发病率和患病率；(d–f) 成年人的死亡数、发病率和患病率；(g–i) 老年人的死亡数、发病率和患病率。

在各州流行病动态中的差异，我们采用K均值聚类，根据各州随时间变化的发病率、患病率和死亡率等流行病学特征对其进行分组。这种聚类方法有助于识别具有相似传播特征和疾病进展模式的州亚组。同一簇中的州具有相近的SIRD 动态，反映出簇内流行病学行为的同质性

各组。纵向轨迹（图1）仅用于探索性分析，以展示各州和年龄组之间的时间异质性，并为聚类分析提供依据。我们采用多种标准选择最优簇数（K），以确保结果稳健并避免过拟合。具体采用以下方法：

肘部法： 将簇内平方和（WCSS） [32] 绘制为簇数的函数，找出 WCSS 下降速度显著放缓的拐点，据此确定合理的簇数。

轮廓系数： 针对每个 K 值计算轮廓系数 [29]，以评估聚类质量。轮廓系数越高，表明簇的定义越清晰、区分度越高。

赤池信息准则和贝叶斯信息准则： 针对每个 K 值计算这些信息准则，以平衡模型拟合度与复杂度。赤池信息准则 [1] 和贝叶斯信息准则 [30] 的值越低，表明模型越优且参数越少。

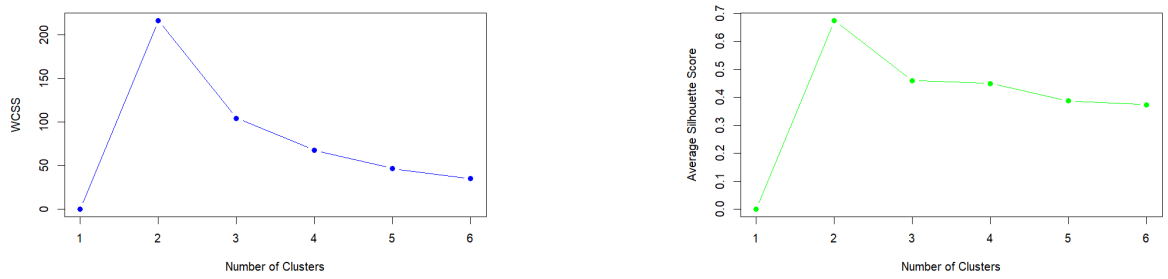
这种聚类方法揭示了水痘传播和康复的不同区域模式，并凸显了同一簇内各州之间有意义的流行病学相似性。

3 数据分析与结果

根据肘部法（图2a），WCSS 在 K$=2$ 和 K $=3$之间显著下降，而在 K $=3$之后下降速度减缓，这表明超过 3 个簇不会显著改善聚类效果。轮廓系数也表明，随着 K $=3$ 的增加，聚类质量有所提高（图2b），这意味着与 K $=2$相比，簇的区分度更高、凝聚性更强。虽然赤池信息准则

和贝叶斯信息准则倾向于选择更简单的模型，即 K = 2（表1），但 K = 3的整体聚类质量和可解释性更优，因此我们选择 K = 3作为最佳簇数。

印度各邦依据水痘流行病学模式分为三个不同的簇：簇1（19个州，占63.4%）包括安得拉邦、阿萨姆邦、比哈尔邦、恰蒂斯加尔邦、德里、古吉拉特邦、哈里亚纳邦、贾坎德邦、卡纳塔克邦、喀拉拉邦、中央邦、马哈拉施特拉邦、奥迪沙邦、旁遮普邦、拉贾斯坦邦、泰米尔纳德邦、特伦甘纳邦、北阿坎德邦和西孟加拉邦；簇——



(a) (b)

图2：（a）肘部法和（b）轮廓系数

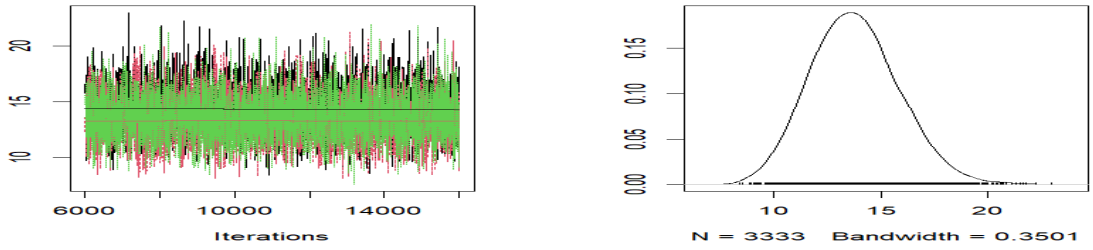
K	轮廓系数	AIC	BIC
2	0.6744202	109.2882	144.3182
3	0.4594910	113.4381	166.6836
4	0.4494454	126.4697	197.9308
5	0.3872293	141.2778	230.9544
6	0.3730322	158.6279	266.5201

表1：确定最佳K值的标准

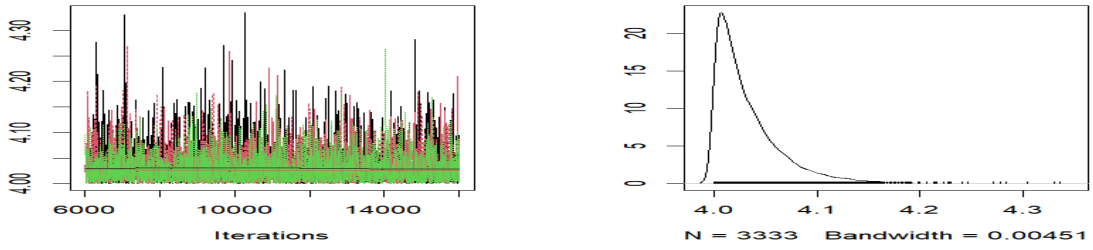
簇2（北方邦，3.3%）仅包含北方邦，而簇3（10个州，33.3%）包括阿鲁纳恰尔邦、果阿邦、喜马偕尔邦、曼尼普尔邦、梅加拉亚邦、米佐拉姆邦、那加兰邦、锡金邦、特里普拉邦，以及德里以外的联邦属地。

3.1 簇特异性发现

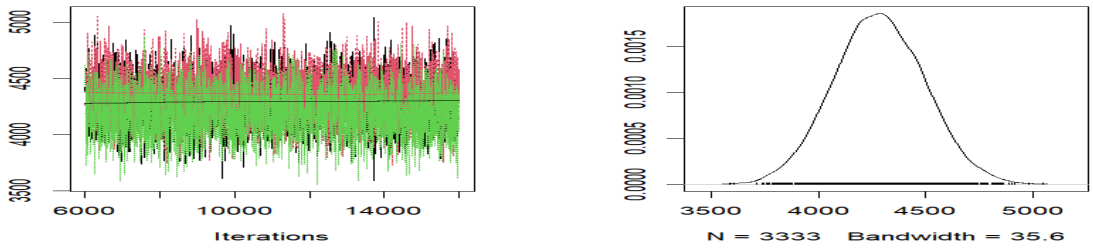
使用 RJAGS 通过马尔可夫链蒙特卡罗获得后验汇总结果，共运行三条链，包含 1,000 次适应迭代、2,000 次预热迭代，以及 10,000 次抽样迭代，稀疏因子为 3。图3展示了精度参数的轨迹图和后验密度图



(a)



(b)



(c)

图3：精度参数 τ_{prev} 在三条马尔可夫链蒙特卡罗链中的轨迹图和后验密度图，分别显示 (a) 簇1、(b) 簇2 和 (c) 簇3。

（离散参数的倒数） τ_{prev} 基于三个簇，并通过三条 MCMC 链进行评估。轨迹图显示链的混合情况良好，围绕稳定均值波动，且没有持续趋势或高自相关等不收敛迹象 [12]。密度图平滑且呈单峰分布，进一步表明后验分布得到了充分探索，估计结果稳定。Gelman-Rubin ˆ 收敛诊断，即该参数的尺度缩减因子（R），簇1为 1.03，簇2为 1.045，簇3为 1.004，均远低于标准收敛阈值 1。1 [11], 这表明所有簇均实现了良好的收敛。结果表明，各链已充分收敛，后验估计稳健可靠。

传播率（β）、康复率（γ）、死亡率（μ）和基本再生数（R_0）的后验均值估计在不同簇和年龄组之间差异显著（表2）。

簇	年龄组	β	γ	μ	R_0
簇1	成年人 (18–59)	6.403	5.553	0.114	1.26
	青少年 (0–17)	7.464	5.644	0.463	1.88
	老年人 (60+)	5.575	3.141	0.831	1.31
簇2	成年人（18–59岁）	2.671	0.596	0.0063	5.42
	青少年（0–17岁）	13.030	9.218	1.569	1.22
	老年人（60+）	13.001	8.055	2.657	1.29
簇3	成年人（18–59岁）	4.570	2.890	0.034	1.55
	青少年（0–17岁）	8.918	7.669	0.505	1.1
	老年人（60+）	6.650	5.550	0.090	1.31

表2：不同年龄组和簇的水痘SIRD参数后验均值。

簇1州代表主要且人口稠密的地区，成年人中的传播率适中（$\beta = 6.40$），青少年中的传播率较高（$\beta = 7.46$）

与成年人和老年人相比（$\beta = 5.58$）。青少年的恢复率最高（$\gamma = 5.64$），成年人次之（$\gamma = 5.55$），而老年人的恢复率较低（$\gamma = 3.14$）。成年人死亡率最低（$\mu = 0.114$），青少年适中（$\mu = 0.463$），老年人最高（$\mu = 0.831$）。这些动态导致基本再生数分别为1.26（成年人）、1.88（青少年）和1.31（老年人），表明青少年是传播的主要推动者，而老年人由于康复较慢、死亡率较高，面临更大的风险。

簇2代表具有极端年龄特异性动态的单个州或地区，成年人中的传播处于中等水平（$\beta = 2.67$），康复（$\gamma = 0.6$）和死亡率（$\mu = 0.0063$）最终得到 $R_0 = 5.42$。青少年（$\beta = 13.03$）和老年人（$\beta = 13.00$）中的传播非常高，康复率适中（分别为$\gamma = 9.22$和8.06），死亡率则较高（分别为$\mu = 1.57$和2.66）。青少年和老年人相应的R_0 值分别为1.22和1.29。这些估计结果凸显出显著的年龄特异性异质性：尽管成年人的传播率低于青少年和老年人，但其暴发潜力仍处于中高水平。

簇3由规模较小的州或东北部各州组成，其青少年中的传播率也较高（$\beta = 8.92$），其次是老年人（$\beta = 6.65$）和成年人（β =4.57）。青少年（$\gamma = 7.67$）和老年人（$\gamma = 5.55$）的恢复率最高，而成年人（$\gamma = 2.89$）的恢复率处于中等水平。各年龄组的死亡率总体较低（μ =0.034–0.505），相应的 R_0 值为1.55（成年人）、1.1（青少年）和1.31（老年人）。这表明，这些规模较小的州具有较低的暴发潜力，同时仍凸显青少年是传播的重要贡献者。

总体而言，簇1州包括人口稠密地区，其青少年和老年人的暴发潜力较高，这是由于康复速度较慢且死亡率较高。

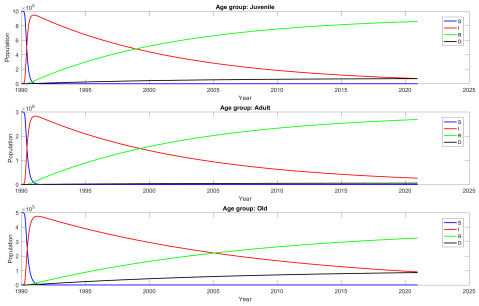
簇2呈现出显著的年龄特异性差异，成年人表现出中等的暴发潜力（$R_0 = 5.42$），而青少年和老年人则维持着很高的传播水平。簇3州在所有年龄组中的传播潜力都较低。这些发现表明，簇1的干预可能需要采取年龄定向策略，例如疫苗接种活动，尤其针对青少年和老年人；而簇3可能主要受益于一般监测和监控。年龄分层分析始终表明，青少年推动传播，成年人维持中等传播水平且死亡率较低，而老年人则因康复速度较慢、死亡率较高而较为脆弱。

3.2 簇基础SIRD动力学的解读

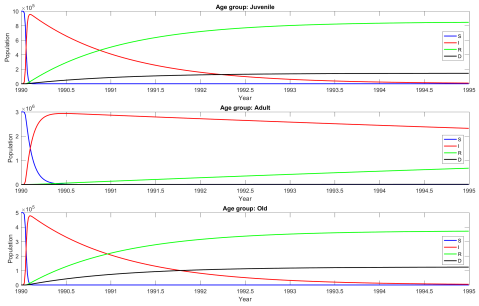
SIRD参数后验估计识别出各簇中三个不同的流行病状态，反映出年龄特异性传播动力学存在显著异质性。尽管控制方程相同，但传播率 β、康复率 γ、死亡率 μ 以及饱和发病函数强度的变化，导致了明显不同的流行病行为。

簇1对应于中等传播情景，传播率、康复率和死亡率较为均衡。如图4(a)所示，感染在早期迅速上升，随后在所有年龄组中平稳下降，并进入逐渐康复阶段。易感人群稳步减少，没有发生骤然崩溃；与此同时，死亡率逐步累积，在青少年和成年人中保持中等水平，而在老年人中更高。饱和效应的存在限制了高患病率下的传播，从而形成相对受控的流行病演变，并伴随持续时间较长的瞬态动力学。

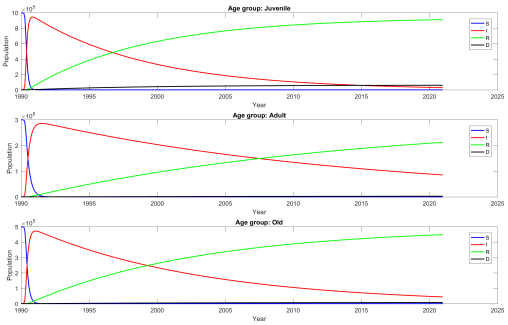
相比之下，簇2呈现出高传播状态（图4(b)）。极强的传播导致青少年和老年人的易感者迅速耗竭，并伴随这些年龄组的感染峰值来得早且十分显著，



(a)



(b)

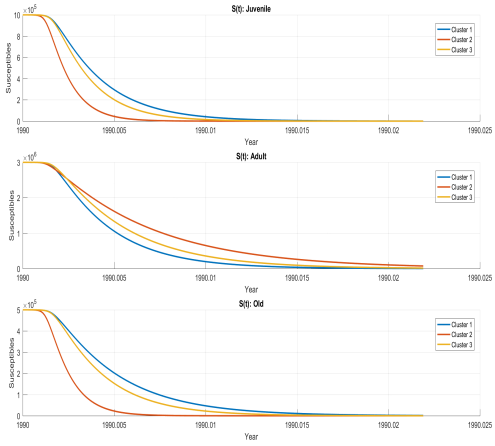


(c)

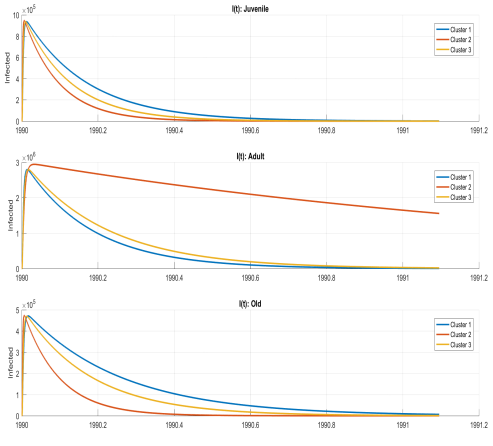
图4：采用表2中的参数值，在中等传播条件下，青少年、成年人和老年人的 SIRD轨迹，基于三个簇；(a) 簇1，(b) 簇2，(c) 簇3。

而在成年人中，感染人群维持在较高水平的时间更长，表明尽管易感者耗竭较慢，传播仍在持续。青少年和老年人的康复者类别迅速增长，而成年康复者则逐渐累积。簇2中的死亡率具有显著的年龄依赖性：青少年和老年人的累计死亡人数较高，而成年人的死亡率相对较低，这是因为疾病导致的成人死亡率较小。

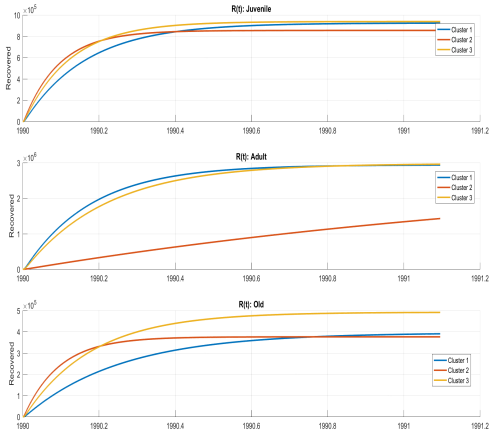
簇3代表介于这两个极端之间的中间情景（图4(c)）。感染峰值处于中等水平，并随时间平稳下降；易感者耗竭



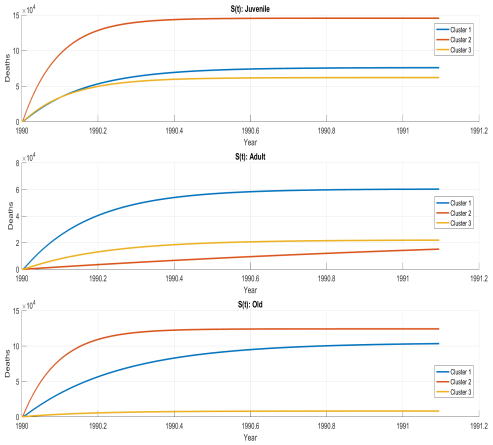
(a)



(b)



(c)



(d)

图5：（a）三个簇中青少年、成年人和老年人的易感者轨迹 $S(t)$，（b）感染者轨迹 $I(t)$，（c）康复者轨迹 $R(t)$，以及（d）死亡轨迹 $D(t)$。

发生得比簇1快，但比簇2慢。更强的饱和效应会在患病率较高时降低有效传播，从而防止失控增长，同时仍允许大范围传播。死亡率始终低于簇1和簇2，尤其是老年人中的死亡率。

年龄分层轨迹为理解这些簇间差异提供了更多见解。图5(a)中的易感者曲线显示，簇2中的青少年易感者消耗最快，其次是簇3；而簇1在更长时间内维持着较大的易感者群体。对于成年人，簇1的易感性下降最快，而簇2的消耗速度相对较慢。老年人的模式再次与青少年相似：簇2的下降最为陡峭，簇1则保留了最大的剩余易感者比例。

图5(b)中的感染者曲线凸显了疫情强度的年龄依赖性差异。对于青少年，簇2的感染峰值最高且出现最早，其次是簇3，而簇1的感染下降较慢。对于成年人，簇2在感染动态中占主导地位，感染人群在较长时间内保持高位；相比之下，簇1和簇3在峰后下降得更快。老年人的感染模式再次与青少年相似：簇2的峰值最为陡峭，而簇1的下降最慢。

图5(c)中的康复者曲线呈现相似的趋势。在青少年和老年人中，簇2的康复者累积最多、速度也最快，簇3的表现居中，而簇1的康复者累积较为缓慢。然而，在成年人中，簇1和簇3形成的康复者群体规模大于簇2，这反映出后者的感染持续时间更长、康复速度更慢。

图5(d)中的死亡轨迹由传播强度与年龄特异性死亡率之间的相互作用所决定。对于青少年，簇2的累计死亡人数最多，其次是簇1，而簇3则较少。对于成年人，簇1的死亡率最高，而尽管簇2的感染负担较高，其死亡率却最低。对于老年人，簇2的死亡曲线最陡、累计死亡人数也最多；簇1

则表现出较高但较低的死亡率，而由于老年人死亡率较低，簇3的死亡率明显更低。

总之，这些结果凸显了参数异质性、年龄结构和发病率饱和在塑造疫情结果方面的关键作用，即使底层模型结构保持不变。

4 模拟研究

为评估所提出的年龄结构化SIRD模型和簇恢复程序的性能，我们在受控条件下开展了模拟研究。研究旨在评估模型能否（i）根据发病率、患病率和死亡轨迹恢复各州的真实簇，以及（ii）准确估计各年龄组的SIRD参数（β, γ, μ）。分析涵盖 $n_{\text{states}} = 10$ 个州 $T = 32$ 年的数据，将其划分为 $n_{\text{clusters}} = 3$ 组，并按年龄分层为成年人、青少年和老年人。各簇和年龄组的真实SIRD参数经过设定，以反映真实的流行病学差异（表3）。

簇	年龄组	β	γ	μ
1	成年人	6.5	5.5	0.1
1	青少年	7.5	5.5	0.5
1	Old	5.5	3.1	0.8
2	成年人	2.4	0.6	0.005
2	青少年	13	9.1	1.5
2	Old	13	8.05	2.6
3	成年人	4.5	2.9	0.03
3	青少年	8.9	7.7	0.5
3	Old	6.6	5.5	0.09

表3：用于模拟的真实值 SIRD 参数。

基于簇的SIRD分舱采用离散时间更新进行模拟：

$$
\begin{aligned}
s_{k,t+1} &= \max(s_{k,t} - \beta_k s_{k,t} i_{k,t}, 0), \\
i_{k,t+1} &= \max(i_{k,t} + \beta_k s_{k,t} i_{k,t} - \gamma_k i_{k,t} - \mu_k i_{k,t}, 0), \\
r_{k,t+1} &= r_{k,t} + \gamma_k i_{k,t}, \\
d_{k,t+1} &= d_{k,t} + \mu_k i_{k,t}.
\end{aligned}
$$

观测数据通过添加真实的测量噪声生成：

$$
\begin{aligned}
y_{k,t}^{\text{inc}} &\sim \text{Poisson}(\text{Inc}_{k,t} \cdot 10^5), \\
y_{k,t}^{\text{prev}} &\sim \text{Normal}(\text{Prev}_{k,t}, 0.01^2), \\
y_{k,t}^{\text{death}} &\sim \text{Poisson}(\text{Death}_{k,t}).
\end{aligned}
$$

随机分配了10个州的簇成员关系，确保每个州都包含全部三个年龄组。

4.1 簇恢复评估

州级特征通过计算发病率、患病率和死亡率随时间变化的均值获得。随后，对这些特征应用K均值聚类，以估计簇。使用调整兰德指数（ARI）评估簇恢复情况：

$$
\text{ARI} = \frac{\text{Index} - \text{Expected Index}}{\text{Max Index} - \text{Expected Index}},
$$

其中，调整兰德指数 $= 1$表示完美康复，调整兰德指数 $= 0$ 表示随机聚类，负值表示康复效果比随机情况更差。

根据模拟研究，真实值簇和估计簇如下：

```
True labels:      3, 3, 3, 2, 3, 2, 2, 2, 3, 1
Estimated labels: 3, 3, 3, 2, 3, 2, 2, 2,1, 3
```

真实簇与估计簇之间的列联表：

真实值 / 估计值	1	2	3
1	0	0	1
2	0	4	0
3	1	0	4

所得调整兰德指数（ARI）为 0.612，表明真实簇分配与估计簇分配高度一致。

4.2 参数恢复

SIRD参数的后验均值估计与真实值高度吻合（表4） ，

证实贝叶斯估计框架能够准确恢复年龄特异性的流行病学参数

簇	年龄组	真实值 β	估计值 β	真实值 γ	估计值 γ	真实值 μ	估计值 μ
1	成年人	6.5	6.410	5.5	4.778	0.1	0.247
1	青少年	7.5	6.275	5.5	5.921	0.5	0.380
1	Old	5.5	4.216	3.1	4.159	0.8	0.524
2	成年人	2.4	4.456	0.6	0.571	0.005	0.0023
2	青少年	13	12.882	9.1	8.780	1.5	0.86
2	Old	13	15.950	8.05	5.220	2.6	1.057
3	成年人	4.5	3.236	2.9	3.859	0.03	0.0841
3	青少年	8.9	6.908	7.7	7.927	0.5	0.102
3	Old	6.6	8.170	5.5	3.053	0.09	0.0443

表4：各簇和年龄组的真实值与估计SIRD参数（β、 γ、 μ）比较。

流行病动态。这些结果证实，所提出的年龄结构化SIRD模型和聚类过程能够可靠地还原流行病动态中潜在的异质性，为其应用于真实GBD水痘数据提供了信心。

5 讨论

本文构建了一个采用饱和型霍林型发病率函数的年龄结构贝叶斯SIRD模型，以研究1990年至2021年印度水痘的长期传播。该模型联合分析了全球疾病负担数据库中的发病率、患病率和死亡率数据，使我们能够在考虑不确定性和漏报的同时，更准确地估计传播率、康复率和死亡率。通过结合年龄分层与邦聚类，该模型捕捉了不同年龄组和地区间疾病动态的重要差异。

结果显示出清晰且一致的流行病学模式。在所有簇中，青少年是传播的主要驱动因素，这可能是因为他们的接触率和易感性较高。成年人通常面临中等程度的传播风险，死亡率较低；而老年人由于康复较慢且疾病相关死亡率较高，死亡风险要高得多。基于簇的分析凸显了显著的地区异质性，表明单一国家级模型无法充分描述印度的水痘传播。这些发现表明，针对不同年龄群体制定公共卫生策略尤为重要，例如为儿童接种疫苗并加强监控，同时为老年人采取防护措施。

从方法学角度来看，霍林型发病率函数允许传播在感染水平较高时趋于饱和，从而避免流行病出现不切实际的快速

增长，同时保持模型的可解释性。贝叶斯框架明确考虑了测量误差和报告偏倚，因此能够获得更可靠的参数估计。模拟研究进一步证实，所提出的方法能够较准确地识别簇并估计年龄特异性模型参数。

应注意本研究的几项局限。首先，分析基于年度数据，难以捕捉短期疫情动态。因此，我们采用了 SIRD 模型的离散时间近似，这种方法适用于长期趋势分析，但不适用于短期预测。其次，分舱模型中的参数可辨识性仍是一项挑战，尤其是在传播率与报告因子密切相关时。尽管使用多数据源有助于缓解这一问题，但仍存在一些不确定性。最后，聚类是通过 K-means 作为事后步骤完成的；这种方法简单且易于解释，但并未完全整合到贝叶斯模型中。

尽管存在这些局限，所提出的框架为研究异质人口中的长期传染病动态提供了一种灵活且可复现的方法。未来的研究可以通过在完整的贝叶斯框架内联合估计簇、纳入疫苗接种覆盖率，或考虑相邻邦之间的空间依赖来扩展该模型。总的来说，本研究表明，结合年龄结构化流行病模型、贝叶斯推断和聚类方法，可以深入了解水痘在印度的不均匀传播，并为制定有针对性的公共卫生策略提供依据。

致谢及声明

作者谨向所有在本研究准备过程中提供宝贵见解和支持的贡献者与合作者表示感谢。

伦理审批声明： 本研究不涉及任何人类参与者或动物实验，因此无需获得正式伦理审批。

经费支持： 本研究未获得公共、商业或非营利部门资助机构提供的任何专项资助。

利益冲突： 作者声明，他们不存在可能影响本文所报告工作的已知竞争性经济利益或个人关系。

数据可用性： 本研究期间生成或分析的所有数据均已包含在本文中。可应要求向通讯作者获取其他数据集。

参考文献

[1] Akaike, H.（1974）。重新审视统计模型识别。*IEEE Transac-tions on Automatic Control*,19(6), 716–723。[2] Allen, L. J. S.（2008）。随机流行病模型导论，数学流行病学，施普林格，81–130。[3] Anderson, R. M., & May, R. M.（1992）。人类传染病：动力学与控制。牛津大学出版社。[4] Bjørnstad, O. N.（2018）。流行病：使用 R 的模型与数据。施普林格。[5] Brauer, F., Castillo-Chavez, C., & Feng, Z.（2017）。流行病学数学模型。施普林格。[6] Chakraborty, C., & Das, K.（2025）。用于含过多零值的多变量纵向结局的潜在类别贝叶斯模型。
https://doi.org/10.48550/arXiv.2509.04804

[7] Chattopadhyay, S., Basu, S., Bhattacharyya, S., Gogoi, M. P., & Das, K.（2025）。多变量二元纵向数据与生存时间数据的同步聚类和联合建模。*Lifetime Data Analysis*, **31**, 830–851。[8] Dietz, K.（1993）。传染病基本再生数的估计。*Statistical Methods in Medical Research*, **2(1)**, 23–41。[9] Diekmann, O., Heesterbeek, J. A. P., & Metz, J. A.（1990）。论基本再生比的定义与计算。*Journal of Mathematical Biology*,**28(4)**:365–382。[10] Fanelli, D. & Piazza, F.（2020）。2019冠状病毒病传播的分析与预测。*Chaos,Solitons & Fractals*,**134**, 109761。[11] Gelman, A. & Rubin, D. B.（1992）。使用多条序列从迭代模拟中进行推断。*Statistical Science*, **7**, 457–511。[12] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B.（2013）。贝叶斯数据分析（第*3*版）。Chapman and Hall/CRC。[13] Genolini, C., Ecochard, R., Benghezal, M., Driss, T., Andrieu, S., & Subtil, F.（2016）。kmlShape：一种依据形状对纵向数据（时间序列）进行聚类的高效方法。*PLOS One*,**11(6)**, e0150738。[14] Hethcote, H. W.（2000）。传染病数学。*SIAM Review*, **42(4)**, 599–653。[15] Holling, C. S.（1959）。捕食与寄生简单类型的若干特征。*Canadian Entomologist*,**91(7)**, 385–398。[16] Ionides, E., Breto, C., & King, A. A.（2006）。非线性动力系统推断。美国国家科学院院刊,**103(49)**, 18438–18443。[17] Kermack, W. O., & McKendrick, A. G.（1927年）。对数学

流行病理论。《英国皇家学会会刊 *A*》，**115**(772)，700–721。[18] Keeling, M. J., & Rohani, P. (2008).人类与动物传染病建模。普林斯顿大学出版社。[19] Kundu, D., Basu, S., Gogoi, M. P., & Das, K. (2025). 一种联合潜在类别贝叶斯模型及其在急性淋巴细胞白血病维持治疗研究中的应用。《应用统计学杂志》。https://doi.org/10.1080/02664763.2025.2511935[20] Misra, V., Gawali, D., Jain, A. K., Khetan, R., & Jain, S. B. (2021). 温度变化导致瓜廖尔市特定地区水痘发病率急剧上升。《印度社区医学杂志》，**46**(2)，323-324。[21] Murray, J. D. (2003).数学生物学 *II*：空间模型与生物医学应用，施普林格，纽约。[22] Muth´en, B., & Shedden, K. (1999). 使用期望最大化算法对混合结局进行有限混合建模。《生物计量学》，**55**(2)，463–469。[23] Nagin, D. S. (1999). 分析发展轨迹：一种半参数的基于群组的方法。《心理学方法》，**4**(2)，139–157。[24] Nagin, D. S., & Odgers, C. L. (2010). 临床研究中的基于群组的轨迹建模。《临床心理学年度评论》，**6**，109–138。[25] Pinto da Costa, J. F., Ferreira, F., Mascarello, M., & Gaio, R. (2023). 使用基于相关性的距离对纵向轨迹进行聚类。《*SN* 计算机科学》，**2**(6)。[26] Plummer, M. (2003). JAGS：用于分析贝叶斯图模型的吉布斯抽样程序。分布式统计计算国际研讨会（*DSC 2003*），维也纳，2003年3月20–22日，1–10。[27] Proust-Lima, C., Philipps, V., & Liquet, B. (2017). 扩展混合模型的估计

使用潜在类别和潜在过程：R 软件包 LCMM。《统计软件杂志》，78(2)，1–56。[28] Roberts, M., & Heesterbeek, M. (2007). 与模型一致的基本再生数估计。《理论生物学杂志》。248(4)，791–799。[29] Rousseeuw, P. J. (1987). 轮廓系数：用于解释和验证聚类分析的图形辅助工具。《计算与应用数学杂志》，20，53–65。[30] Schwarz, G. (1978). 模型维数的估计。《统计年刊》，6(2)，461–464。[31] Singh, M. P., Singh, G., Kumar, A., Singh, A., & Ratho, R. K. (2011). 流行病学启示：印度北部昌迪加尔周边农村社区的水痘暴发调查。《印度病理学与微生物学杂志》，54(4)，772–774。[32] Thorndike, R. L. (1953). 谁属于这个家族？《心理测量学》，18(4)，267–276。[33] Vaidya, S. R., Tilavat, S. M., Kumbhar, N. S., & Kamble, M. B. (2018). 印度达德拉-纳加尔哈维利联邦属地部落与工业区的水痘暴发。《流行病学与感染》，146(4)，476–480。[34]van den Driessche, P., & Watmough, J. (2002). 传染病传播分舱模型的再生数和次阈值平衡点。《数学生物科学》，180，29–48。[35] Vazquez, M., LaRussa, P., Gershon, A., Steinberg, S., & Brown, L. (2004). 水痘流行病学。《临床微生物学评论》，17(3)，383–401。[36] Zhou, J., Zhang, Y., & Tu, W. (2023). clusterMLD：一种高效的多变量纵向数据层次聚类方法。《计算与图形统计学杂志》，32(3)，1131–1144。