arXiv:2602.06148v1 [stat.AP] 5 Feb 2026

版本日期：2026年2月9日

种群动态的非线性驱动因素：非参数溯祖方法

Filippo Monti[1], Nuno R. Faria[2], Xiang Ji[3], Philippe Lemey[4], Moritz U.G. Kraemer[5,6], Marc A. Suchard[1,7,8]

[1] 生物统计学系，乔纳森和卡琳·菲尔丁公共卫生学院，加州大学洛杉矶分校，美国加利福尼亚州洛杉矶[2] 传染病流行病学系，公共卫生学院，帝国理工学院，英国伦敦[3] 统计学系，爱荷华州立大学，美国爱荷华州埃姆斯[4] 微生物学、免疫学和移植学系，雷加研究所，鲁汶大学，比利时鲁汶[5] 生物学系，牛津大学，英国牛津[6] 大流行科学研究所，牛津大学，英国牛津[7] 生物数学系，大卫·格芬医学院（加州大学洛杉矶分校），加州大学洛杉矶分校，美国加利福尼亚州洛杉矶[8] 人类遗传学系，大卫·格芬医学院（加州大学洛杉矶分校），加州大学洛杉矶分校，美国加利福尼亚州洛杉矶

通讯作者：Marc A. Suchard，生物统计学、生物数学和人类遗传学系，加州大学洛杉矶分校，695 Charles E. Young大道南段，洛杉矶，加利福尼亚州 90095-7088，美国；电子邮箱： msuchard@ucla.edu

摘要 有效群体大小（N_e（t））是群体遗传学和系统发育动力学中的一个基本参数，可量化遗传多样性并揭示人口史。基于溯祖的方法能够根据分子序列数据重建的时间标度系统发育树，推断 N_e（t）随时间变化的轨迹。要理解种群动态的生态和环境驱动因素，需将 N_e（t）与气候或流行病学变量等外部数据联系起来。现有方法通常假设协变量与 N_e（t）之间存在对数线性关系，这可能无法捕捉复杂的生物过程；当真实关系为非线性时，还可能引入偏差。我们提出一种灵活的贝叶斯框架，通过高斯过程（GP）先验，将协变量纳入采用分段常数 N_e（t）的溯祖模型中。GP 是定义在函数空间上的分布，由核函数及数据驱动的超参数控制；它无需严格的参数假设，即可自然地适应非线性协变量效应。这一建模方法能更准确地估计协变量与N_e（t）之间的关系，在关联呈非线性时减轻偏差，并提供可解释的不确定性量化，且其不确定性会随协变量空间而变化。为兼顾协变量驱动的全局模式与局部时间动态，我们将 GP 先验与高斯马尔可夫随机场结合，以约束 N_e（t）轨迹的平滑性。通过在高维潜变量场上运行哈密顿蒙特卡罗，实现了高效推断。通过模拟研究和三项实证应用——巴西黄热病毒动态（2016–2018）、晚第四纪麝牛人口动态，以及喀麦隆 HIV-1 CRF02 AG 的演化——我们证明，该方法既能在适当情况下验证线性关系，也能揭示以往可能被忽略或误判的非线性协变量效应。该框架推进了系统动力学推断，使研究者能够更准确、更符合生物学实际地模拟环境与流行病学因素如何随时间影响种群大小。

关键词：贝叶斯系统发育动力学；合并理论；高斯过程；哈密顿蒙特卡洛；有效群体大小

1 引言

共祖过程该**合并模型**由Kingman (1982年年提出，用于描述中性进化条件下遗传谱系样本的谱系过程。自提出以来，人们已发展出许多扩展模型，以适应更复杂的人口统计与进化情景。

有效群体大小该模型的一个关键参数是**有效群体大小**(N_e)，它表示一个理想化种群的大小；在该种群中，遗传漂变程度与所研究的实际种群相同。普查群体大小统计所有个体，而有效群体大小则考虑繁殖成功率的差异、世代重叠和种群大小波动等因素；这些因素会影响向过去追溯时谱系的合并速率（Hudson，1990年）。在系统发育学和群体遗传学中，根据遗传数据估算有效群体大小，对于理解过去种群动态、选择压力和人口史至关重要。

时变有效群体大小最初的共祖模型（Kingman，1982年）假设种群大小恒定，这种简化往往并不现实。为克服这一局限，Griffiths 和 Tavar´e (1994) 扩展了该模型，使其允许有效群体大小随时间变化，即 $N_e = N_e(t)$。这一推广推动了大量研究的发展，旨在构建推断方法以重建有效群体大小的动态变化。

参数模型 对时变有效群体大小进行建模的一种方法是采用参数模型，即假设种群大小遵循具有少量参数的预定义函数形式。一个常见的例子是指数

增长模型，正如 Kuhner 等人(1998年) 和 Drummond 等人 (2002年) 所采用的那样。参数模型的优势在于，它们能够以紧凑且易于解释的方式表示人口史。然而，如果所选择的函数形式与潜在趋势不匹配，那么模型设定错误问题可能导致推断出现显著偏差。因此，要获得准确结果，就需要针对多种函数形式耗费大量时间进行模型选择过程。这促使了半参数和非参数方法的发展，以实现更大的灵活性。

分段常数模型 Pybus 等人 （2000年）通过将函数 $N_e(t)$ 建模为分段常数，开启了新的研究方向。具体而言，作者允许 $N_e(t)$ 在每个合并时间发生变化，即每当发现一个最近共同祖先时都发生变化（经典天际线图估计）。这意味着自由参数的数量等于观测值的数量；换言之，每个有效群体大小估计值都根据单个观测值推断得出，因此估计结果具有很大的变异性且噪声较大。

为解决这些局限性，人们提出了多种方案。 Strimmer 和 Pybus （2001年）通过采用基于赤池信息准则（AIC）校正的模型选择方法，减少了上述模型中的自由参数数量（广义经典天际线图估计）。在贝叶斯框架下，Drummond 等人（2005）提出预先指定变点总数上限。为提高灵活性，Opgen-Rhein 等人 （2005）借助可逆跳跃马尔可夫链蒙特卡罗算法（Green，1995）为该数量指定了先验。然而，结果对先验选择高度敏感。

Minin 等人 (2008年提出了一种替代方案：像 Pybus 等人 (2000年) 那样，将变点锚定在共祖时间上，但为

种群大小函数的分段常数水平指定了一个（一阶）高斯马尔可夫随机场（GMRF）先验。他们的主要思路是，通过施加相邻水平之间的依赖关系，采用一种惩罚 $N_e(t)$ 在短时间窗口内变异性的先验来减少过拟合（贝叶斯天际线）。随后，Gill 等人 (2013) 采用了同样的基于高斯马尔可夫随机场的方法，但预先固定了 一个等间距的变点网格（贝叶斯*Skygrid*）。这种方法克服了先验对树结构的依赖，而这种依赖此前限制了该方法在涉及多棵树的分析中的应用。在此基础上，Gill 等人 (2016年) 通过对有效群体大小（对数）水平的高斯马尔可夫随机场均值采用对数线性模型，引入了外部协变量。这使得可以引入额外信息，以进一步辅助推断过程，在非相邻水平之间建立依赖关系；同时，作为副产物，也揭示了哪些因素可能影响人口波动。由于$N_e(t$水平可能数量众多，且常呈现较强的后验相关性，传统的随机游走马尔可夫链蒙特卡洛方法往往混合缓慢。因此，Baele 等人 (2020年) 采用一种基于梯度的采样器——哈密顿蒙特卡罗（HMC）——更高效地模拟其后验分布。

非参数模型作为分段常数模型的替代方案，Palacios 和 Minin(2013) 提出了一种完全非参数的方法，将有效群体大小 $N_e(t$建模为时间的连续函数。他们为一个时间潜在函数设置了高斯过程（GP）先验，并通过逻辑变换将其与种群大小的倒数关联起来，从而确保 $1/N_e(t)$ 保持为正且有界。这种变换使他们能够将共祖过程视为点过程，并使用基于稀疏化的数据增广方案计算其似然；该方案以一组有限的模拟潜在点取代难以处理的时间积分。尽管这一策略为 $N_e(t)$ 提供了平滑、灵活的轨迹，但也带来了大量计算开销，因为每次 MCMC 迭代都必须重新采样这些

潜在点。此外，该框架完全以时间作为唯一预测变量，因此若不重新定义稀疏化过程和似然结构，就难以纳入其他协变量。

现有模型的局限性 分段常数模型仍是最受欢迎的选择，因为它们兼顾了灵活性与计算可行性。其主要局限在于，难以在 $N_e(t)$ 级别之间保持平滑性，同时又保留足够多的变点，以捕捉有意义的人口变异。Minin 等人Minin etal.(2008年) 和 Gillet al. (2013) 提出的高斯马尔可夫随机场先验在一定程度上解决了这一问题，即关联相邻的 $N_e(t)$ 级别。 Gill 等人(2016年) 将这一思路扩展到纳入外部协变量，从而建立了超越相邻级别的依赖关系。然而，他们的对数线性形式将 $N_e(t)$ 与协变量之间的关系限制在狭窄的函数类中；当这一假设不成立时，推断会产生偏差，也无法充分利用协变量提供的全部信息。

提出的模型 我们提出一种混合方法，在分段常数模型的计算效率与非参数方法的灵活性之间取得平衡。在 Gill 等人 (2016年) 的基础上，我们保留了具有固定变点网格的高斯马尔可夫随机场先验，但将对数线性均值结构替换为定义在协变量上的高斯过程先验。通过为广泛的一类函数赋予正概率，GP 先验使我们的模型能够捕捉协变量的非线性效应，并解决对数线性模型的局限性。此外，GP 先验由参数化的 核函数 刻画，该核函数控制随机函数在不同输入点处取值之间的协方差。核函数及其参数的选择决定了推断函数的平滑性。这不仅使 $N_e(t)$ 的非相邻水平之间能够存在高阶依赖关系，还可以通过为核函数参数赋予先验，实现数据驱动的平滑性选择

。与 Palacios 和 Minin (2013) 使用基于时间的高斯过程，并将其与 $N_e(t)$ 的归一化变换相结合，且依赖删减采样来计算合并似然不同，我们的模型：(1) 使用高斯马尔可夫随机场先验保持时间连贯性；(2) 使用高斯过程整合协变量信息，以刻画非线性效应；(3) 避免基于删减采样的推断所带来的计算负担。

论文结构 以下各节将首先正式介绍一般的时变合并模型（2.1），然后将其调整为有效种群大小函数采用分段常数形式建模的框架（2.2），并区分同期取样和序列取样。随后，第2.3 节将探讨将共祖过程用作先验时需要考虑的问题。接着，我们介绍对 $N_e(t)$ 水平所采用的先验（2.4），计算后验分布的梯度和海森矩阵（2.5.1），并说明在基于HMC的采样方案中使用它们的理由（2.5.2）。最后，为检验模型的可靠性及其在现实世界中的适用性，我们将所提方法应用于合成数据和真实数据示例（3）。

2 方法

概述与约定。在贝叶斯系统发育学中，**合并过程** 用作树先验，即一种概率模型，用于描述对取样得到的遗传谱系沿时间逆向追溯时它们如何合并。给定分子序列数据，该先验将人口统计参数（即有效群体大小 N_e）与谱系树的形状及分支长度联系起来。

我们用 $\mathcal{F}$ 表示一棵有根二叉**系统发育树**，描述 K 个采样分类单元的祖先关系。由于 $\mathcal{F}$ 是二叉树，它有 K 个末端节点（叶节点）和 $K-1$ 个内部节点，每个内部节点都代表一次共祖事件，即两条谱系合并为一个共同祖先。系统发育树 $\mathcal{F}$ 由两个部分完全刻画：（i）一个拓扑结构，用于指定

哪些谱系发生合并；（ii）一组节点时间，包括叶节点时间和内部节点（共祖）时间。

在本文中，我们采用逆向时间 约定：时间从现在向过去递增，因此数值越小表示时间越近。为简化记号，我们首先假设等时取样，即所有 K 叶节点均在当前时刻同时取样。在此假设下，我们可以将所有叶节点时间设为零， $t_0 = 0$，并将内部节点时间记为

$$0 < t_1 \leq t_2 \leq \cdots \leq t_{K-1}, \tag{1}$$

其中 t_k 表示第 k次合并事件的逆向时间。每个内部节点代表在该时刻存在的两条谱系的最近共同祖先（MRCA）。在第2.2节中，我们将推广该设定以纳入序列取样（异时取样），即叶节点可能在不同时间取样。

在给定人口统计参数的条件下，合并模型为有序合并时间集合（t_1，...， t_{K-1}）指定一个与 $\mathcal{F}$相容的概率密度，同时我们假设树拓扑服从均匀分布。第 2.3节通过扩展合并过程，明确考虑随时间合并的谱系身份，从而构造带标签的树上的概率分布。

2.1 一般时变模型

现在，我们在有效种群大小可能随时间变化的假设下，将共祖过程形式化，即 $N_e = N_e(t)$。在等时取样下，有序序列 $t_1, \ldots, t_{K-1}$ 列出了**共祖时间**（按回溯时间计算），其中t_k 是第 k次合并事件发生的时间。令 n_k 表示现存谱系的数量

在区间 $[t_{k-1}, t_k)$ 内；当在 t_k 处发生一次合并事件时，我们有递推关系 $n_k = n_{k-1} - 1 = K + 1 - k$。

固定时变有效群体大小函数 $N_e(\cdot)$后，第 k个 t_k 共祖事件在 t_{k-1} 时发生的条件概率密度为（GriffithsandTavar´e，1994）：

$$p(t_k \,|\, t_{k-1}, N_e(\cdot)) = \binom{n_k}{2} \frac{1}{N_e(t_k)} \exp\left\{ -\binom{n_k}{2} \int_{t_{k-1}}^{t_k} \frac{1}{N_e(\tau)}\, d\tau \right\}. \tag{2}$$

这里，$\binom{n_k}{2} N_e(t_k)^{-1}$ 是瞬时概率，也就是风险率，表示 n_k个节点中任意两个在时间 t_k 合并的概率，而指数项则是在（t_{k-1}，t_k）期间不发生合并的生存概率。对各事件求积可得共祖时间的联合密度：

$$p(t_1, \ldots, t_{K-1} \,|\, N_e(\cdot)) = \left\{ \prod_{k=1}^{K-1} \binom{n_k}{2} \frac{1}{N_e(t_k)} \right\} \exp\left\{ -\sum_{k=1}^{K-1} \binom{n_k}{2} \int_{t_{k-1}}^{t_k} \frac{1}{N_e(\tau)}\, d\tau \right\}. \tag{3}$$

2.2 分段常数模型

一种常见的用于建模时变有效群体大小函数 $N_e(t)$ 的半参数方法，是假设该函数在固定的时间区间网格上呈**分段常数**。令 t_0，t_1，……，t_{G-1} 表示一系列**网格点**，将时间轴划分为从 $t_0 = t_0$开始的 G 个区间。这些网格点被视为外生的，意味着它们的设定独立于实际实现的共祖时间。一个简单且有效的选择是使用等间距网格，不过也可以采用任何单调序列。于是，有效群体大小被定义为

$$N_e(t) = \theta_g \quad \text{for } \bar{t}_{g-1} < t \le \bar{t}_g, \qquad g = 1, 2, \ldots, G-1. \tag{4}$$

由于根时间可能超出最后一个预先定义的网格点，我们在其后追加一个位于足够久远过去的点 t_G ，并将 t_{G-1} 与 t_G 之间的种群大小记为 θ_G。该区间的长度不具有推断作用；在实践中，我们只需令最后一个网格区间延伸至实际实现的根高度即可。

等时数据 为便于说明，我们首先在假设所有 K 分类单元均于同一时间采样（等时情形）的条件下推导联合密度。令 c_g 表示网格点 $\bar{t}_{g-1}$ 与 $\bar{t}_g$之间的共祖事件（内部节点）数。此外，令

$$\mathbf{t}_g = \left(t_{g,1}, t_{g,2}, \ldots, t_{g,c_g}\right) \tag{5}$$

表示时间 t_{g-1} 与 t_g之间共祖事件的序列。我们进一步将这些项汇总为向量 $\mathbf{t} = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_G)$。最后，令 $n_{g,k}$ 表示时间 $t_{g,k}$之前紧邻时刻的未共祖节点数。

按照第 2.1 节相同的风险率–生存推理，所有共祖时间的联合概率密度可写为

$$\begin{aligned} p(\mathbf{t} \mid \boldsymbol{\theta}) &= \left\{ \prod_{g=1}^{G} \prod_{k=1}^{c_g} \left[\binom{n_{g,k}}{2} \frac{1}{\theta_g} \right] \right\} \exp\left\{ -\sum_{g=1}^{G} \sum_{k=1}^{c_g} \binom{n_{g,k}}{2} \frac{t_{g,k} - t_{g,k-1}}{\theta_g} \right\} \\ &= \left\{ \prod_{g=1}^{G} \left[\frac{1}{\theta_g}\right]^{c_g} \prod_{k=1}^{c_g} \binom{n_{g,k}}{2} \right\} \exp\left\{ -\sum_{g=1}^{G} \frac{1}{\theta_g} \sum_{k=1}^{c_g} \binom{n_{g,k}}{2} \left(t_{g,k} - t_{g,k-1}\right) \right\}, \end{aligned} \tag{6}$$

其中 $\binom{0}{2} = 0$ ，为方便记述，我们定义边界值 $t_{g,\,0} := t_{g-1}$ 和 $t_{g,\,c_g+1} := t_g$，这样指数中的内层求和便可写成统一形式。

异时数据 为了考虑末端采样时间的异质性，我们需要重新定义 $t_{g,\ k}$，使其可以是一个合并时间 或一个末端采样日期，并相应更新 $n_{g,\ k}$。事实上，末端采样日期的顺序会影响在特定时间窗口内能够发生合并的谱系数量。例如，假设 $t_{g,\ k}$ 与某个采样日期重合（而非合并事件），那么后续发生合并事件的概率就会增加，因为更多节点可以拥有共同祖先。

我们用 c_g 来统计 $(\bar{t}_{g-1}, \bar{t}_g]$ 之间发生的共祖事件数量，而用d_g 表示同一时期内末端采样日期的数量。此外，我们引入集合 $\{\mathcal{C}_g\}$，其中 $\mathcal{C}_g$ 收集了在¯ ¯网格点 t_{g-1} 和 t_g 之间发生的所有共祖事件的索引。根据这一新记法，似然变为:

$$p(\mathbf{t} \mid \boldsymbol{\theta}) = \left\{ \prod_{g=1}^{G} \left[\frac{1}{\theta_g} \right]^{c_g} \prod_{k \in \mathcal{C}_g} \binom{n_{g,k}}{2} \right\} \exp\left\{ -\sum_{g=1}^{G} \frac{1}{\theta_g} \sum_{k=1}^{c_g+d_g} \binom{n_{g,k}}{2} (t_{g,k} - t_{g,k-1}) \right\}. \quad (7)$$

唯一值得注意的区别是，指数中的内部求和变得更加细化：如上所述，这是为了考虑到采集新样本后共祖事件概率的增加。

2.3 基于共祖的似然

方程（7）中给出的共祖过程表述为 何时发生事件（即内部节点时间）指定了一个概率分布，但对哪些谱系发生合并并未作出限定。因此，它定义了一个关于具有相同有序共祖时间的树的等价类的密度。若要获得完全标记的树上的概率分布，必须扩展此密度，以纳入每次合并所涉及谱系的身份。等价地，对于给定的有序共祖时间序列，其概率质量在所有与这些时间相容的带标签的树之间均匀分布。

为了计算此类树的数量，我们固定一组有序共祖时间以及每个事件发生前一刻相应的现存谱系数量。

在每个共祖时间 $t_{g,\ k}$，$n_{g,\ k}$ 现存谱系中的任意一对无序谱系都可能合并，从而产生 $\binom{n_{g,\ k}}{2}$ 种可能的选择。由于每种选择都唯一确定了后续的谱系构型，因此，与给定有序共祖时间相容的不同带标签的树的总数为

$$\prod_{g=1}^{G} \prod_{k \in \mathcal{C}_g} \binom{n_{g,k}}{2}. \tag{8}$$

将（7）除以该数量，即可在带标签的树上生成经过适当归一化的密度。

多基因座推广 如果存在更多谱系 $\mathcal{F}_1$，$\ldots \mathcal{F}_F$（例如与基因组中的不同基因座相关联的谱系），并且假设在给定有效群体大小的条件下它们相互独立，那么可以将它们的联合分布写成如下形式：

$$p\Big(\{\mathcal{F}_i\}_{i=1}^{F} \,|\, N_e(\cdot)\Big) = \prod_{i=1}^{F} p(\mathcal{F}_i \,|\, N_e(\cdot))\,. \tag{9}$$

此后介绍的所有方法都可以轻松推广，以容纳额外的谱系。

2.4 种群大小建模

为避免每个有效群体大小的正值约束所带来的复杂性，我们将重点放在其对数变换上：

$$\gamma_g = \log \theta_g. \tag{10}$$

随后，我们利用两类信息对水平序列 γ_g 进行建模：一是时间顺序，遵循这样的设想，即我们不预期 $N_e(t)$ 会在短时间内发生突变；

（2）外部协变量，它们为推断过程提供更多信息，有助于限制突变，并揭示哪些因素可能驱动人口波动。

在第 2.4.1 节中，我们重点研究如何基于高斯过程先验直接利用协变量对 γ_g 进行建模。接着，在第 2.4.2 节中，我们利用时间顺序信息，使用（一阶）高斯马尔可夫随机场先验对高斯过程误差进行建模；换言之，我们使用高斯过程对高斯马尔可夫随机场均值进行建模。这种联合方法兼顾了协变量所带来的跨所有层级的全局依赖性，以及局部时间依赖性。

2.4.1 全局协变量依赖：高斯过程（GP）先验

假设在每个网格点 g，我们观测到一组 P协变量 $\mathbf{x}_g = (x_{1g}, x_{2g} \ldots, x_{Pg})$；例如，$x_{ig}$ 可以是时间 g的温度。这些协变量的选择主要有两个目的：一是指导推断过程，例如，当我们先验地知道某个变量与病原体的人口史相关时；二是直接考察它们与 $N_e(t)$ 的关联。随后，我们可以将对数有效群体大小设为这些协变量的随机（实值）函数加上一个误差项，如下所示：

$$\gamma_g = \log \theta_g = f(\mathbf{x}_g) + \epsilon_g, \tag{11}$$

其中，误差 ϵ_g 假设服从均值0 和精度矩阵 $\mathbf{P}^*$的多元正态分布，即 $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots\ldots, \epsilon_G) \sim \mathcal{MVN}(\mathbf{0}, (\mathbf{P}^*)^{-1})$。

继Minin 等人（2008年）之后，Gill 等人（2013）假设了一个常数函数f（$\mathbf{x}_g$）$\equiv \beta_0$，并将精度$\mathbf{P}^*$ 限制为（一阶）高斯马尔可夫随机场先验所特有的结构。我们将在下一节讨论并说明这一选择的动机。Gill 等人（2016年）后来扩展了该框架，假设采用（对数）线性模型来描述

协变量，即将协变量函数固定为 $f(\mathbf{x}_g) = \sum_{i=1}^{P} x_{ig}\beta_i$，并对线性系数 β_i赋予独立同分布正态先验。尽管便于解释，对数线性假设严重限制了候选函数形式的范围。除非愿意对一系列协变量变换进行模型选择，否则这一局限既会妨碍充分利用协变量信息，也会导致无法捕捉非线性关系。

为克服这一限制，我们在共祖函数与协变量之间的关系上设定**高斯过程**（**高斯过程**）先验，从而大幅增强模型捕捉复杂非线性依赖关系的能力。为便于表述，我们假设协变量效应以加性方式组合，即$f(\mathbf{x}_g) = \sum_{i=1}^{P} f_i(x_{ig})$（尽管如下文所述，这并非必要条件）。接着，我们可以为每个函数 $f_i()$ 设定相互独立的高斯过程先验，使得

$$\gamma_g = \log\theta_g = \sum_{i=1}^{P} f_i(x_{ig}) + \epsilon_g \quad \text{with} \quad f_i(\cdot) \overset{\text{ind}}{\sim} \mathcal{GP}\left(0\ ,\ k_{\boldsymbol{\alpha}_i}(\cdot,\cdot)\right), \tag{12}$$

其中，$k_{\boldsymbol{\alpha}_i}(\cdot,\ \cdot)$ 是高斯过程核，用于衡量每对协变量点处函数值的协方差，而 $\boldsymbol{\alpha}_i$ 是其超参数。

高斯过程背景知识 高斯过程（GP）是一种灵活的未知函数先验 f。它可以看作是多元正态分布（MVN）的无限维推广；更重要的是，它的定义保证了对于任意有限组 m 输入 x_1、.....、x_m，相应的函数值（$f(x_1)$.....、$f(x_m)$）服从联合高斯分布。高斯过程由两个部分完全定义：均值函数$\mu(\cdot)$和协方差函数（核函数） $k(\cdot,\ \cdot)$。均值函数描述每个输入处函数值的期望，而核函数决定不同点处函数值之间的协方差。根据这些定义，向量（$f(x_1)$,

.....、 $f(x_m)$)服从多元正态分布，其均值为（$\mu(x_1)$.....、$\mu(x_m)$），协方差矩阵的第 ij 个元素为 $k(x_i, x_j)$。

高斯过程先验在不假设特定函数形式的情况下，为建模非线性关系提供了极大的灵活性。高斯过程模型的一个关键要素是核函数的选择。实际上，核函数衡量输入之间的相似性，并决定从该过程中抽取的函数的平滑性和形状。例如，平方指数核倾向于生成更平滑的函数，而 Mat´érn 核则更能适应变化更不规则的情况（Rasmussen 和 Williams，2006）。

核函数选择 在我们的情境中，我们预期函数整体较为平滑，最多只有两到三个拐点。实际上，协变量的影响不太可能多次改变符号。因此，平方指数核似乎是最合适的选择。该核函数由两个正的超参数定义：边际尺度（σ^2）和长度尺度（ℓ），其定义如下：

$$k(x_i, x_j) = \sigma^2 \exp\left\{-\frac{(x_i - x_j)^2}{2\ell^2}\right\}. \tag{13}$$

边际尺度控制函数变化的整体幅度，而长度尺度决定正则性程度，或相关性随输入空间中的距离增加而衰减的速度。具体而言， ℓ取值较低时，函数振荡更多；增大 ℓ 则会使函数趋于线性。在贝叶斯框架下，我们可以为每个超参数指定先验分布，从而根据数据确定其取值。我们建议为两个核函数超参数分别采用速率为 1 的独立指数先验。此外，如果高斯过程明显过拟合，也就是说，考虑到先验生物学知识，其振荡速度显得不切实际，我们建议通过设定严格大于零的下界来约束长度尺度，从而抑制过短程的相关性。

边际分布 回到方程（12），我们可以利用先验分布与误差相互独立这一点，写为：

$$\gamma_g \sim \mathcal{N}\left(\mathbf{0}\,,\; k_{\boldsymbol{\alpha}}(\mathbf{x}_g, \mathbf{x}'_g) + (\mathbf{P}^*)^{-1}_{gg}\right), \tag{14}$$

其中 $k_{\boldsymbol{\alpha}}(\mathbf{x}_g,\ \mathbf{x}'_g) := \sum_{i=1}^{P} k_{\boldsymbol{\alpha}_i}(x_{ig},\ x'_{ig})$。用向量形式表示时，我们可以汇集所有对数有效种群大小水平 $\boldsymbol{\gamma} = (\gamma_1,\ \gamma_2,\ \ldots,\ \gamma_G)$，并将其联合分布表示为多元正态分布：

$$\boldsymbol{\gamma} \sim \mathcal{MVN}\left(\mathbf{0}\,,\; \boldsymbol{K}_{\boldsymbol{\alpha}}(\mathbf{X}\,, \mathbf{X}') + (\mathbf{P}^*)^{-1}\right), \tag{15}$$

其中 $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2,\ \ldots,\ \mathbf{x}_G)$ 是一个矩阵，其各列对应网格点特定协变量，而 $\boldsymbol{K}_{\boldsymbol{\alpha}}(\mathbf{X},\ \mathbf{X}')$ 是一个矩阵，其元素 $(g_i,\ g_j)$ 等于 $k_{\boldsymbol{\alpha}}(\mathbf{x}_{g_i},\ \mathbf{x}'_{g_j})$。这一框架可扩展为纳入协变量之间的交互作用，从而实现更加灵活的推断。

2.4.2 建模误差精度：局部时间依赖与高斯马尔可夫随机场

误差精度矩阵 $\mathbf{P}^*$ 为误差结构选择合适的形式 $\boldsymbol{\epsilon}$的精度矩阵 $\mathbf{P}^*$ 至关重要。事实上，选择完全不受限制的矩阵 $\mathbf{P}^*$ 会引入 $G(G+1)/2$ 个额外参数，大幅增加参数空间维数，并可能导致不可识别性问题或过拟合。一种常见的解决方案是将 $\mathbf{P}^*$ 限制为对角矩阵，也就是说，假设高斯误差ϵ_g 相互独立，且方差可能相同。然而，这种方法存在若干缺点。首先，模型会对协变量和（“观测到的”）函数值中的离群值非常敏感。例如，与协变量离群值相关的 $N_e(t)$ 水平会具有更高的变异性，因为 GP 没有足够的信息来捕捉其取值。当协变量离群值出现在时间范围的两端时，这一问题会进一步加剧

在时间范围两端观测到时。在 GP 推断中，离群的观测到的GP函数值会显著影响后验函数，进而造成偏差，尤其是在使用平方指数核等平滑核函数时。尽管在我们的情境中，GP函数值是 $\log N_e(t)$ 水平，并未被直接观测到，但其后验分布受观测数据通过合并先验的影响，因此会出现类似但较轻微的问题。

其次，误差相互独立意味着，在给定协变量的条件下，相邻网格点之间不存在残差相关性。这意味着模型无法解释协变量未能说明的任何局部时间依赖。然而，由于网格点按时间顺序排列，我们预期 $N_e(t)$ 相邻时间点之间存在残差相关性，这表明矩阵 $\mathbf{P}^*$ 应具有更具结构性的形式。

高斯马尔可夫随机场 遵循 Minin 等人(2008年) 和Gill 等人(2013) 的方法，我们使用一阶 高斯马尔可夫随机场对误差进行建模，使相邻误差之间产生直接的共同依赖。具体而言，我们假设向量 $\boldsymbol{\epsilon}$服从均值为 $\mathbf{0}$ 、精度为 $\mathbf{P}^* = \tau\mathbf{P}$ 的多元正态分布，其满足：

$$p(\boldsymbol{\epsilon}) \propto \exp\left\{-\frac{\tau}{2}\boldsymbol{\epsilon}'\mathbf{P}\boldsymbol{\epsilon}\right\} = \exp\left\{-\frac{\tau}{2}\sum_{g=2}^{G}(\epsilon_g - \epsilon_{g-1})^2\right\}, \tag{16}$$

其中， τ 是标量精度参数，$\mathbf{P}$ 是一个三对角矩阵，其对角元素等于$\mathbf{P}_{ii} = 2$ （ $i = 2,\ldots, G-1$ ）以及 $\mathbf{P}_{1,1} = 1 = \mathbf{P}_{G,G}$，非对角元素等于 $\mathbf{P}_{i,i-1} = \mathbf{P}_{i,i+1} = -1$。$\mathbf{P}$ 的结构决定了这种方法的主要计算优势：高斯马尔可夫随机场先验可在线性时间内计算。

精度参数 τ 在控制高斯马尔可夫随机场的平滑性方面发挥着关键作用。

τ 的值越大，连续值之间的平方差（$\epsilon_g - \epsilon_{g-1}$）^{2}, 受到的惩罚就越重。这会迫使这些值彼此更加接近，从而得到更平滑的误差序列。相反， τ的值越小，连续值之间就允许出现更大的跳跃，从而形成平滑度较低、或更加锯齿状的序列。因此，精度参数充当控制局部时间依赖程度的超参数，我们可以直接根据数据估计其值。

精度参数的超先验 我们建议为高斯马尔可夫随机场的精度参数指定形状参数为 1、尺度参数为 10 的逆伽马先验。这一选择取代了以往研究中采用的过于弥散的替代方案，例如形状参数和尺度参数分别为 0.001 和 1000 的伽马先验。事实上，我们发现，在我们的情境下，伽马 先验在零附近集中了过多的概率质量，导致边际方差不切实际地偏大。采用的设定提供了更均衡的折衷：既保留较弱的信息性，又能避免对潜在过程进行病态的欠平滑。

另一种高斯马尔可夫随机场参数化 Minin 等人提出的方法 （2008年）和Gill 等人（2013），随后由 Gill 等人（2016年）扩展以纳入协变量，采用了一种不同但等效的参数化方式，为求完整，我们也在此报告。具体而言，他们的“以高斯马尔可夫随机场为中心”的方法直接使用高斯马尔可夫随机场对水平建模，其均值设定为协变量的随机函数，即 γ_g $\mu_g = \mu_g$（$\mathbf{x}^g$）。Gill 等人（2016年）将此类函数限制为（对数）线性形式；而在本文中，如上所述，我们为其指定了更灵活的高斯过程先验。完整的层级模型随后可重写为：

archical 模型可重写为:

$$\boldsymbol{\gamma}|\mathbf{X} \sim \mathcal{MVN}\left(\boldsymbol{\mu}\left(\mathbf{X}\right), \frac{1}{\tau}\mathbf{P}^{-1}\right) \tag{17}$$

$$\boldsymbol{\mu}\left(\mathbf{X}\right) \sim \mathcal{MVN}\left(\mathbf{0}\ ,\ \boldsymbol{K}_{\boldsymbol{\alpha}}(\mathbf{X}\ , \mathbf{X}')\right). \tag{18}$$

其中 $\boldsymbol{\mu}\left(\mathbf{X}\right) := (\mu_1(\mathbf{x}_1)\ \ \mu_2(\mathbf{x}_2) \ldots, \mu_G(\mathbf{x}_G))$ 是高斯马尔可夫随机场均值向量。

2.5 推断

在本研究中，我们采用贝叶斯框架，推断种群的进化历史及其随时间发生的人口变化。系统发育学的核心目标是重建系统发育树（$\mathcal{F}$），以展示遗传序列（$\mathbf{y}$）之间的进化关系。通常假设一个由速率矩阵（$\mathbf{Q}$）参数化的连续时间马尔可夫链（CTMC），用于模拟 DNA 替换，并通过沿树模拟进化来生成这些序列。典型的系统发育模型还假设进化速率可能随时间或位点而变化。本研究不影响这些建模选择，因此为简化符号，不会显式地对进化速率进行参数化。

贝叶斯推断的基础是贝叶斯定理，它使我们能够根据观测数据（$\mathcal{F}$）计算未知参数的后验分布，例如系统发育树（$\mathcal{F}$）和速率矩阵（$\mathbf{Q}$）（$\mathbf{y}$）。该后验分布与数据似然和未知参数的先验分布之积成正比。在系统发育学中，数据似然 $p(\mathbf{y}\,|\,\mathcal{F},\ \mathbf{Q})$ 等于观测到的序列由 CTMC 沿树生成的概率。对 $\mathbf{Q}$ 的先验选择不在本文讨论范围内，并取决于 CTMC 速率矩阵的参数化方式（例如，参见 Kimura，1980年；Jukes 和 Cantor，1969年）。与本文互补的、关于 $\mathbf{Q}$ 先验的讨论——也就是说，

us结合协变量的高斯过程见 Monti 等人 (2025) .

相反，我们在此关注树先验。具体而言，我们采用一个合并过程先验 $p(\mathcal{F}|\, N_e(t))$，其中种群大小为分段常数（见方程 (7)）。接着，我们结合高斯马尔可夫随机场和高斯过程先验（如第 2.4 节所述），利用时间信息和外部协变量，为种群大小水平 $p(N_e(t)|\mathbf{X}$ 指定先验。现在，我们可以综合这些要素，写出给定观测数据$\mathbf{y}$ 时未知参数的后验分布 $(\mathcal{F}, N_e(t),\mathbf{Q})$：

$$p(\mathcal{F}, N_e(t), \mathbf{Q} \,|\, \mathbf{y}) \propto p(\mathbf{y} \,|\, \mathcal{F}, \mathbf{Q})\, p(\mathbf{Q})\, p(\mathcal{F}|\, N_e(t))\, p(N_e(t) \,|\, \mathbf{X})\,. \tag{19}$$

由于后验分布较为复杂，无法直接计算，因此我们需要采用模拟方法从中抽取样本。常见的选择是使用马尔可夫链蒙特卡罗（MCMC）。传统的 MCMC 方法（例如随机游走马尔可夫链蒙特卡罗）长期以来一直是对树和速率参数进行抽样的标准方法，但在对大量高度相关的参数进行抽样时往往效率低下（Neal，2011；Gelman 等，2013），而分段常数种群大小通常正是这类参数。

Baele 等人(2020年) 提出，在 Gill 等人(2016年) 引入的对数线性模型框架下，采用哈密顿蒙特卡罗（HMC）采样器，以提高对有效种群大小水平进行采样的效率。HMC 是一种基于梯度的马尔可夫链蒙特卡罗采样器，特别适用于高维且高度相关的后验分布（Neal，2011）。通过利用梯度信息引导采样过程，HMC 能够在参数空间中提出更大且更高效的移动，从而克服随机游走方法的局限性。下面我们将展示如何轻松扩展他们的框架，以从我们提出的基于高斯过程的模型的后验分布中进行采样。在接下来的章节中，我们将首先推导 HMC

采样器的两个关键要素：合并似然和先验的梯度与海森矩阵。随后，我们将详细解释 HMC算法如何利用这些量，高效地从后验分布中进行采样。

2.5.1 梯度与海森矩阵的计算

我们首先推导共祖似然 $p(\mathcal{F} \mid N_e(t))$ 的对数及其对应先验 $p(N_e(t) \mid \mathbf{X})$ 的梯度和海森矩阵。

共祖似然 如第 2.3 节所述，在异时数据框架下，基于合并过程的对数似然可写为共祖时间分布与一个归一化常数之和，即：

$$\log p(\mathcal{F} \mid N_e(t)) = \log p(\mathbf{t} \mid \boldsymbol{\gamma}) + \text{const} \tag{20}$$

$$= -\sum_{g=1}^{G} c_g \gamma_g - \sum_{g=1}^{G} e^{-\gamma_g} \sum_{k=1}^{c_g+d_g} \binom{n_{g,k}}{2} (t_{g,k} - t_{g,k-1}), \tag{21}$$

在第二个等式中，我们代入了公式 (7)，将 $\boldsymbol{\theta}$ 替换为对数有效群体大小水平 $\boldsymbol{\gamma}$，并注意到常数项相互抵消。

应用简单的求导法则，我们可以求出关于每个参数 γ_g 的梯度，即向量 $\nabla_{\boldsymbol{\gamma}}$log $p(\mathbf{t} \mid \boldsymbol{\gamma})$，其第 g 个分量为：

$$\frac{\partial \log p(\mathbf{t} \mid \boldsymbol{\gamma})}{\partial \gamma_g} = -c_g + e^{-\gamma_g} \sum_{k=1}^{c_g+d_g} \binom{n_{g,k}}{2} (t_{g,k} - t_{g,k-1}). \tag{22}$$

Besi因此，海森矩阵 $\nabla^2_{\boldsymbol{\gamma}}$log $p(\mathbf{t} \mid \boldsymbol{\gamma})$ 是一个对角矩阵，其对角元素为 s:

$$\frac{\partial^2 \log p(\mathbf{t} \mid \boldsymbol{\gamma})}{\partial \gamma_g^2} = -e^{-\gamma_g} \sum_{k=1}^{c_g+d_g} \binom{n_{g,k}}{2} (t_{g,k} - t_{g,k-1}). \tag{23}$$

先验梯度和海森矩阵根据方程（17），我们可以将 $\boldsymbol{\gamma}$ 在给定均值向量 $\boldsymbol{\mu}$（$\mathbf{X}$）的条件下的对数先验写为：

$$\log p(\boldsymbol{\gamma} \mid \boldsymbol{\mu}(\mathbf{X}), \tau) = -\frac{\tau}{2} [\boldsymbol{\gamma} - \boldsymbol{\mu}(\mathbf{X})]^{\top} \mathbf{P} [\boldsymbol{\gamma} - \boldsymbol{\mu}(\mathbf{X})] + \text{const.} \tag{24}$$

因此，相对于 $\boldsymbol{\gamma}$ 的梯度和（负定）海森矩阵为

$$\nabla_{\boldsymbol{\gamma}} \log p(\boldsymbol{\gamma} \mid \boldsymbol{\mu}(\mathbf{X}), \tau) = -\tau \mathbf{P} [\boldsymbol{\gamma} - \boldsymbol{\mu}(\mathbf{X})], \tag{25}$$

$$\nabla^2_{\boldsymbol{\gamma}} \log p(\boldsymbol{\gamma} \mid \boldsymbol{\mu}(\mathbf{X}), \tau) = -\tau \mathbf{P}. \tag{26}$$

后验梯度和海森矩阵。后验梯度 $\nabla_{\boldsymbol{\gamma}} \log p(\boldsymbol{\gamma} \mid \mathbf{t}, \mathbf{X})$ 和海森矩阵 $\nabla^2_{\boldsymbol{\gamma}} \log p(\boldsymbol{\gamma} \mid \mathbf{t}, \mathbf{X})$ 直接由相应先验和似然贡献之和得到。

2.5.2 使用哈密顿蒙特卡罗进行采样

针对共祖过程似然和高斯过程先验，存在计算上可行的精确梯度，这使得我们能够在贝叶斯框架下使用 HMC采样器进行高效推断（Neal，2011）。给定感兴趣的参数 $\boldsymbol{\theta}$，其后验分布为π（$\boldsymbol{\theta}$），**HMC采样器** 通过引入一个辅助变量扩展采样空间，并模拟哈密顿动力学来探索这个扩展后的空间，从而生成 Metropolis-Hastings 提议（Metropolis 等人，1953）。用动力学的术语来说，感兴趣的参数 $\boldsymbol{\theta}$ 称为系统的“位置”分量，而辅助变量$\mathbf{v}$ 则解释为“动量”或“速度”。 $\boldsymbol{\theta}$ 和$\mathbf{v}$ 被假定为相互独立；前者的分布是目标后验分布 π（$\boldsymbol{\theta}$），而后者通常建模为协方差矩阵为$\mathbf{M}$的中心化多元正态随机变量。

该系统由一个总能量函数 $H(\boldsymbol{\theta}, \mathbf{v})$ 全局描述，该函数称为哈密顿函数，定义为系统“势能” $U(\boldsymbol{\theta})$ 与“动能” $K(\mathbf{v})$ 之和。进而，势能 $U(\boldsymbol{\theta})$ 设为对数后验密度 $U(\boldsymbol{\theta}) = -\log \pi(\boldsymbol{\theta})$ 的负值，而动能函数等于 $K(\mathbf{v}) = \mathbf{v}'\mathbf{M}^{-1}\mathbf{v}$。从当前状态 $(\boldsymbol{\theta}_0, \mathbf{v}_0)$ 出发，HMC 通过在空间 $(\boldsymbol{\theta}, \mathbf{v})$ 中模拟哈密顿动力学，生成一个 MH 提议（Metropolis 等人，1953年）；该动力学由以下微分方程支配：

$$\begin{aligned} \frac{d\boldsymbol{\theta}}{dt} &= \quad \nabla K(\mathbf{v}) = \mathbf{M}^{-1}\mathbf{v} \\ \frac{d\mathbf{v}}{dt} &= -\nabla U(\boldsymbol{\theta}) = \nabla \log \pi(\boldsymbol{\theta}). \end{aligned} \tag{27}$$

一种成功的数值近似这些动力学的方法基于蛙跳算法 (Neal，2011），它通过大小为 ϵ的离散步长向前积分这些动力学，具体如下：

$$\begin{aligned} \mathbf{v}_{t+\frac{\epsilon}{2}} &= \mathbf{v}_t + \frac{\epsilon}{2}\nabla \log \pi(\boldsymbol{\theta}_t) \\ \boldsymbol{\theta}_{t+\epsilon} &= \boldsymbol{\theta}_t + \epsilon \mathbf{M}^{-1}\mathbf{v}_{t+\frac{\epsilon}{2}} \\ \mathbf{v}_{t+\epsilon} &= \mathbf{v}_{t+\frac{\epsilon}{2}} + \frac{\epsilon}{2}\nabla \log \pi(\boldsymbol{\theta}_{t+\epsilon}). \end{aligned} \tag{28}$$

将哈密顿量模拟一段时间 t 总共需要 t/ϵ 步，而这又需要大约相同次数的梯度计算。然而，哈密顿蒙特卡罗提议往往与当前状态相当不相关，并以高概率被接受（Neal，2011）。这些特性使该算法在从高维相关分布中抽样时特别有效（Beskos等人，2013；Gelman等人，2013）。

预条件 协方差矩阵M的一个常见选择是单位矩阵。一个主要动机是，这种方法可避免在每个蛙跳步中进行代价高昂的矩阵求逆（对于稠密无结构矩阵，时间复杂度为立方）。然而，通过利用后验分布的几何结构并采用“知情”质量矩阵，HMC采样器的效率可以大幅提升（Neal，2011）：这一技术称为**预条件**。一种候选做法是将质量矩阵设为后验海森矩阵的负值（Girolami和Calderhead，2011）。幸运的是，在当前情形下，海森矩阵是三对角矩阵，因此使用专门的高斯消元算法计算其逆与动量的乘积$\mathbf{M}^{-1}\mathbf{v}$，是一项线性时间复杂度的运算。一种更快的替代方案是将所有非对角元素直接设为零。这意味着，在每个蛙跳步中，唯一与预条件相关的运算是用各海森矩阵对角元素的倒数对动量变量进行缩放。

3 结果

在本节中，我们使用两个合成数据示例和三个真实数据示例来验证模型。具体而言，我们将首先模拟并分析两棵树，其中将有效群体大小的对数设定为协变量的线性函数和非线性函数。其次，我们根据2016年至2018年间收集的样本，研究温度波动对巴西黄热病毒（YFV）人口史的影响（Hill等人，2022年），以及温度波动对全球晚第四纪麝牛的影响（Campos等人，2010；Gill 等人，2016年）。随后，我们利用 HIV 的局部发病率，追踪1996年至2004年喀麦隆 HIV-1 CRF02 AG支系的种群动态（Brennan等人，2008年；Faria et al.，2012）。以下所有分析均使用

公开提供的软件 BEAST X（Baele et al.，2025），并由高性能计算环境 BEAGLE（Ayres et al.，2019年）提供支持。用于复现分析的代码和数据可在公共代码库中获取（https://github.com/suchard-group/NonparametricCoalescentProcesses）。

3.1 模拟

为了评估所提出的方法在估计模型参数方面的性能，我们开展了两项模拟研究。我们的参考数据基于Hill等人（2022年）的研究结果。他们利用705个本地样本，研究了2016年至2018年巴西各地平均气温与黄热病毒传播之间的关系，以及该病毒的人口波动。随后，我们在分段常数合并模型下模拟了两棵系统发育树，每棵树均包含705个分类单元。在这些模拟中，我们分别将对数有效群体大小设为关于Hill等人（2022年）所提供的（标准化）平均温度值的线性函数和凹函数。这两个函数在图1的两个图中以虚线表示。

随后，我们使用以温度为协变量的高斯过程先验分析这两棵模拟树。我们采用平方指数核，并对其两个超参数设置速率为 1 的指数先验。图 1 中的两个图展示了中位数 $\log N_e(t)$（实线）随协变量变化的情况，以及相应的 95% 最高后验密度（HPD）区间，分别对应线性情形（左图）和凹形情形（右图）。推断得到的 $\log N_e(t)$ 与真实值高度吻合，证明了我们方法的有效性和准确性。

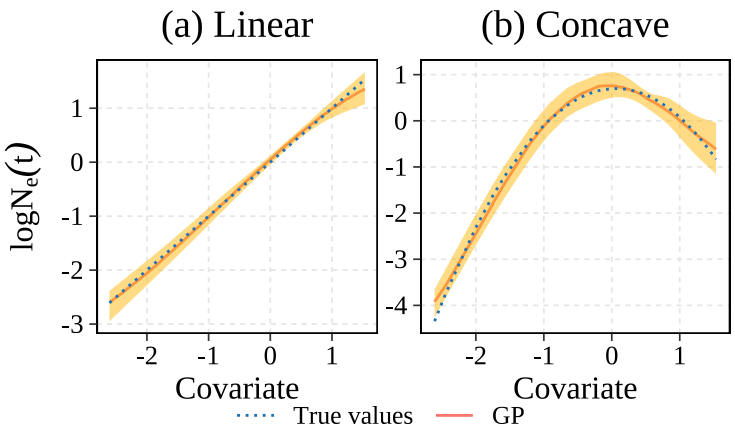

FIGURE 1: SIMULATIONS: $\log N_e(t)$ VS COVARIATES

注释：这两幅图展示了两种模拟情景下，有效群体大小的对数与协变量之间的关系。所选协变量为2014年至2019年巴西的平均气温（Hill等人，2022年）。在分段常数合并模型下模拟了两棵树，其中对数 $N_e(t)$水平分别设为协变量的线性函数（左图）和非线性函数（右图），如两条虚线所示。实线表示高斯过程模型对这两个函数的中位数估计，阴影区域表示95%最高后验密度区间。

3.2 温度对巴西黄热病毒人口动态的影响 il

在两个模拟数据集上验证了我们的方法之后，我们将其应用于真实数据。我们再次从 Hill等人(2022年) 的研究着手；该研究分析了 2014年 至 2018年 巴西东南部黄热病毒（YFV）的气候与人口动态之间的关系。研究者还发现，YFV 种群大小与气温水平之间存在正相关关系，这表明较高的气温有利于病毒的繁殖和传播。为此，研究者将 YFV 的分段常数有效群体大小建模为巴西东南部当地平均气温的对数线性函数。在此，我们旨在放宽这一严格的对数线性假设，以更好地指导推断过程，并检测温度影响中可能存在的非线性关系。为此，我们采用与 Hill等人 (2022年) 一致的系统发育分析，使用 705 条近乎——

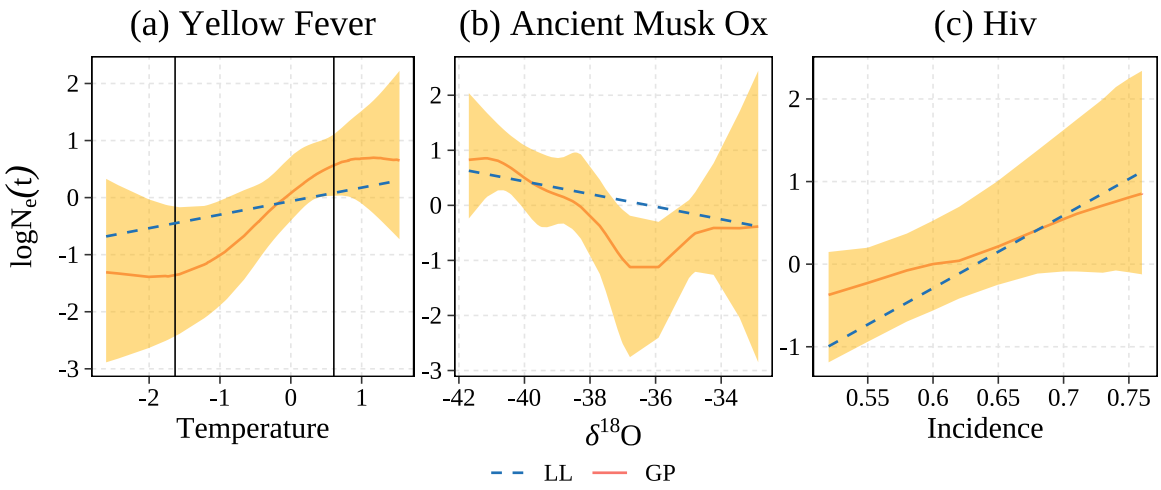

FIGURE 2: DATA EXAMPLES: $\log N_e(t)$ VS COVARIATES

注：这三个图分别使用对数线性（LL）模型和高斯过程（GP）模型，比较对数有效群体大小与三个协变量——温度、 $\delta^{18}O$ （逆温度的代用指标）以及（HIV）发病率——之间的关系。三个数据示例分别聚焦于：(a) 巴西的黄热病毒（2016—2018年），数据来自Hill等人(2022年)；(b) 古代麝牛（Campos等人，2010；Gill 等人，2016年）；(c) 喀麦隆的 HIV CRF02 AG毒株（1990 – 2004）（Brennan等人，2008年；Gill 等人，2016年）。在推断过程中，假设进化树是随机的，并使用分段常数 $N_e(t$的共祖先验对其建模。实线表示高斯过程模型下，对数 $N_e(t$水平关于协变量的中位数估计，阴影区域表示相应的95%最高后验密度区间。虚线表示LL模型下的中位对数 $N_e(t_\circ$ 最后，第一幅图中的黑色竖线表示“趋平”点，即温度效应改变符号或变为零的位置。

在人类、新热带灵长类和蚊子中采样的完整YFV基因组。不过，我们使用以温度为协变量的灵活高斯过程先验，对分段常数有效群体大小进行建模。与第3.1节一样，我们采用平方指数核函数，并对其两个超参数分别赋予率为1的指数先验。

图2(a)分别展示了采用对数线性模型（虚线）和高斯过程（实线）模型得到的中位数对数 $N_e(t$关于气温水平的函数。基于高斯过程的函数揭示了种群大小与温度之间存在S形非线性关系。具体而言，我们识别出两个“趋平”点，以图2(a)中的黑色竖线表示：当温度介于这两个点之间时，温度升高与种群增长相关；相反，当温度处于这一区间之外时，温度升高对种群没有影响或影响可以忽略不计（导数小于0.05）。由于对数线性模型假定温度效应恒定，因此它会对非线性函数进行插值，从而分别低估和高估温度在中间范围内和该范围之外升高所产生的影响。

图3 底部的两幅折线图比较了温度与 $N_e(t)$随时间变化的趋势。垂直阴影带分别对应低于和高于前述趋平点的温度。仔细观察可发现，当温度处于两个趋平点之间的范围内（白色区域）时，温度与 $N_e(t)$往往同步变化；而在该范围之外，两者看起来相当独立。

3.3 晚第四纪麝牛种群动态

晚第四纪大型哺乳动物的衰退和灭绝已有充分记录，但其原因仍存在争议（Lorenzen等人，2011）。讨论主要围绕气候变化还是人类活动才是首要因素展开

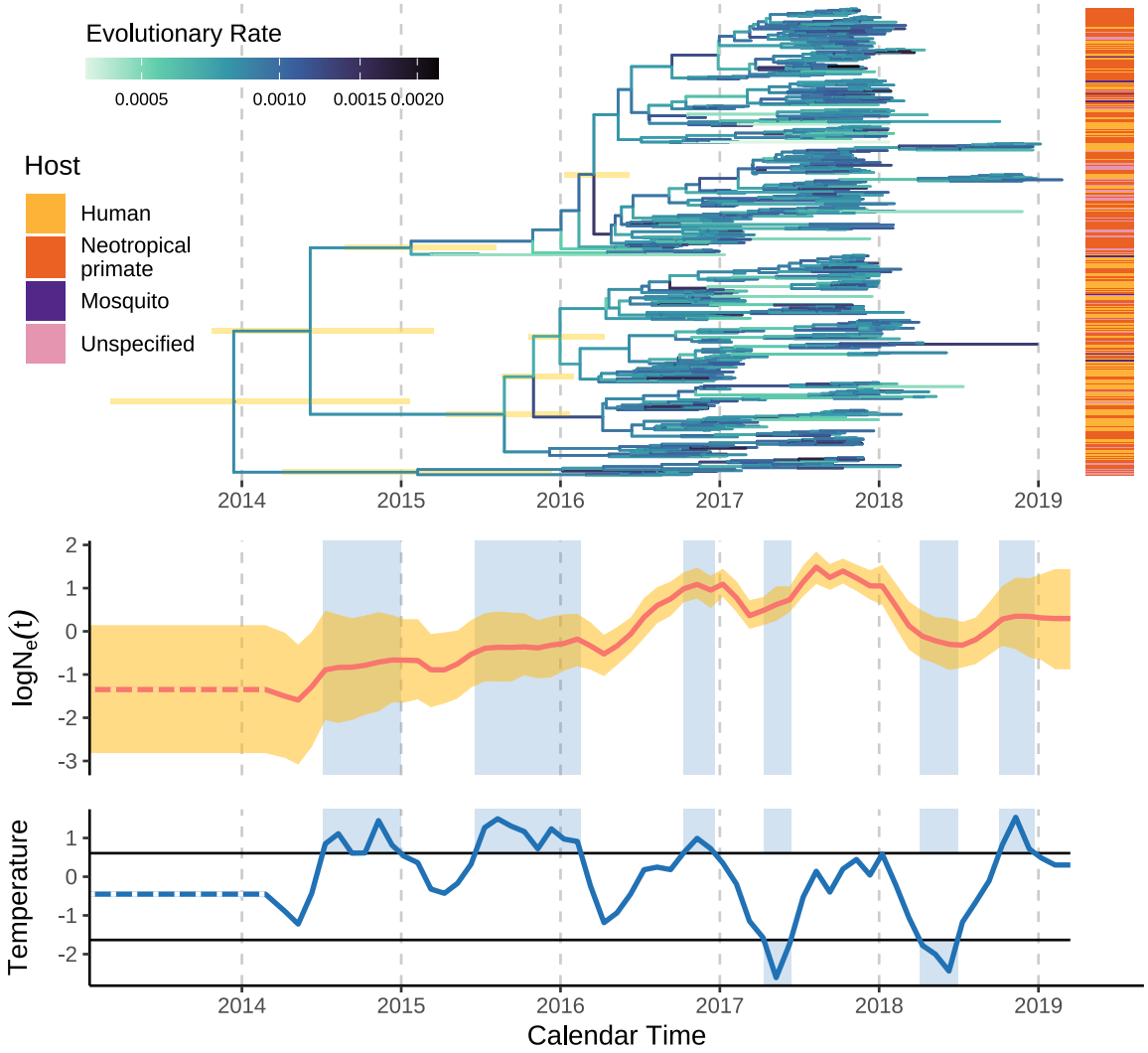

FIGURE 3: YELLOW FEVER VIRUS: TREE AND DEMOGRAPHY

注：图的上半部分展示了根据2016年至2018年在巴西采集的人类、新热带灵长类、蚊子及其他宿主的样本构建的黄热病毒（YFV）进化树；右侧竖直矩形的颜色标明了各样本所属类别（Hill等人，2022年）。黄色阴影横条表示部分节点年龄的95%最高后验密度区间，分支则根据推断出的分支恒定DNA替换进化速率着色。图的下半部分两张折线图展示了2014年至2019年间巴西YFV的log $N_e(t)$水平和当地平均气温的趋势。红线表示在高斯过程模型下推断出的log $N_e(t)$中位数，黄色阴影区域表示相应的95%最高后验密度区间。蓝色阴影竖带表示温度分别高于或低于此处两条黑色水平线所示的两个“趋平”点的时间段；在图2(a)中，这两个点以两条竖线表示。在白色区域内，气温升高与YFV种群增长相关；而在阴影区域内，气温升高则与种群稳定相关。假定2014年以前气温保持恒定（虚线），且等于Hill等人(2022年)报告的最早可用数值。

（Lorenzen等人，2011）。研究人员经常利用古代遗传物质重建种群历史，随后可将其与气候及化石记录进行比较，从而为这一争论提供启示。

一项2010年由Campos等人（2010）开展的研究利用古DNA重建了麝牛的种群历史，所依据的是贝叶斯Skygrid（Minin 等人，2008年）和Skyline（Drummond 等人，2005）模型。他们发现，麝牛种群似乎并未受到人类进入其栖息地的影响。然而，该研究确实发现种群增长与较冷气候期之间存在相关性，种群在更温暖、更不稳定的时期有所下降。这表明，环境变化而非人类才是麝牛种群变化的主要原因。为进一步探究这一问题，Gill 等人（2016年）用一个基于温度代用指标的对数线性模型扩展了Skygrid，该代用指标通常用于重建古气候趋势，即氧同位素记录。作为氧同位素组成的度量，他们采用了冰芯 $\delta^{18}O$ 数据，这些数据来自格陵兰冰芯计划（Dansgaard等人，1993；Grootes等人，1993），并按3000年的时间间隔取平均值，其中较高的 $\delta^{18}O$ 水平对应于更温暖的极地温度。作者推断的种群动态历史表明，气候与 $\delta^{18}O$ 水平之间存在反向关系，从而证实了Campos等人（2010）的研究结果。然而，对数线性模型中负的协变量效应大小并未揭示这两个变量之间存在统计学上显著的关联。

在此，我们探究对数线性假设是否过于严格，无法刻画这种关系。我们使用与Campos等人（2010）和Gill 等人（2016年）相同的数据，重现了系统发育分析。具体而言，我们使用了序列数据，这些数据由从149个标本中采集的线粒体控制区682 bp序列组成；这些标本的年代从现在至放射性碳年距今56,900年不等。样本采集自泰梅尔半岛（$n = 54$）、乌拉尔（$n = 26$）、东北

西伯利亚（$n = 12$）、北美洲（$n = 14$）和格陵兰岛（$n = 43$）。

采用使用平方指数核函数、超参数先验为速率 1 的指数分布的高斯过程模型，我们还检测到 $\delta^{18}O$ 与麝牛对数有效群体大小之间存在反向关系（图2(b)）。我们观察到，在较低温度下（$\delta^{18}O < -37$），变暖的影响似乎大于 LL 模型所显示的程度，也就是说，相应的种群规模收缩更大。此外，在较高温度下（$\delta^{18}O \geq -37$），效应量似乎接近 0，但存在相当大的不确定性，这很可能是样本量有限所致。图4 展示了 $N_e(t)$（红线）和 $\delta^{18}O$ （蓝线）随时间变化的趋势。我们绘制 $\delta^{18}O$ 时采用了递减坐标轴，以突出相关性；因此，两条线总体上相互平行，意味着温度与 $N_e(t)$ 呈反向关系。相反，我们注意到，在较高温度下，这种关系看起来即使不是正相关，也至少趋于正相关，尽管较宽的 HPD 区间表明仍存在相当大的不确定性。值得注意的是，基于GP的方法细致地量化了温度效应的不确定性如何随不同气温水平而变化，这支持了这样一种观点：变暖气候在较低温度下的影响更为确定，而在较高温度下则存在显著更大的不确定性。

最后，在图4的人口动态曲线上方，我们展示了古代麝牛的最大支系可信度（MCC）树，以描述其进化历史，树梢根据样本采集地点着色。我们识别出两个主要支系，它们大致将采自格陵兰岛和加拿大的样本与采自泰梅尔半岛和乌拉尔的样本区分开来，这与先前的研究一致。

3.4 喀麦隆HIV-1 CRF02 AG支系的人口史

重组HIV-1基因组（称为循环重组型，CRF）是在两种或多种不同的HIV-1亚型结合时形成的。CRF02 AG是一种特定的

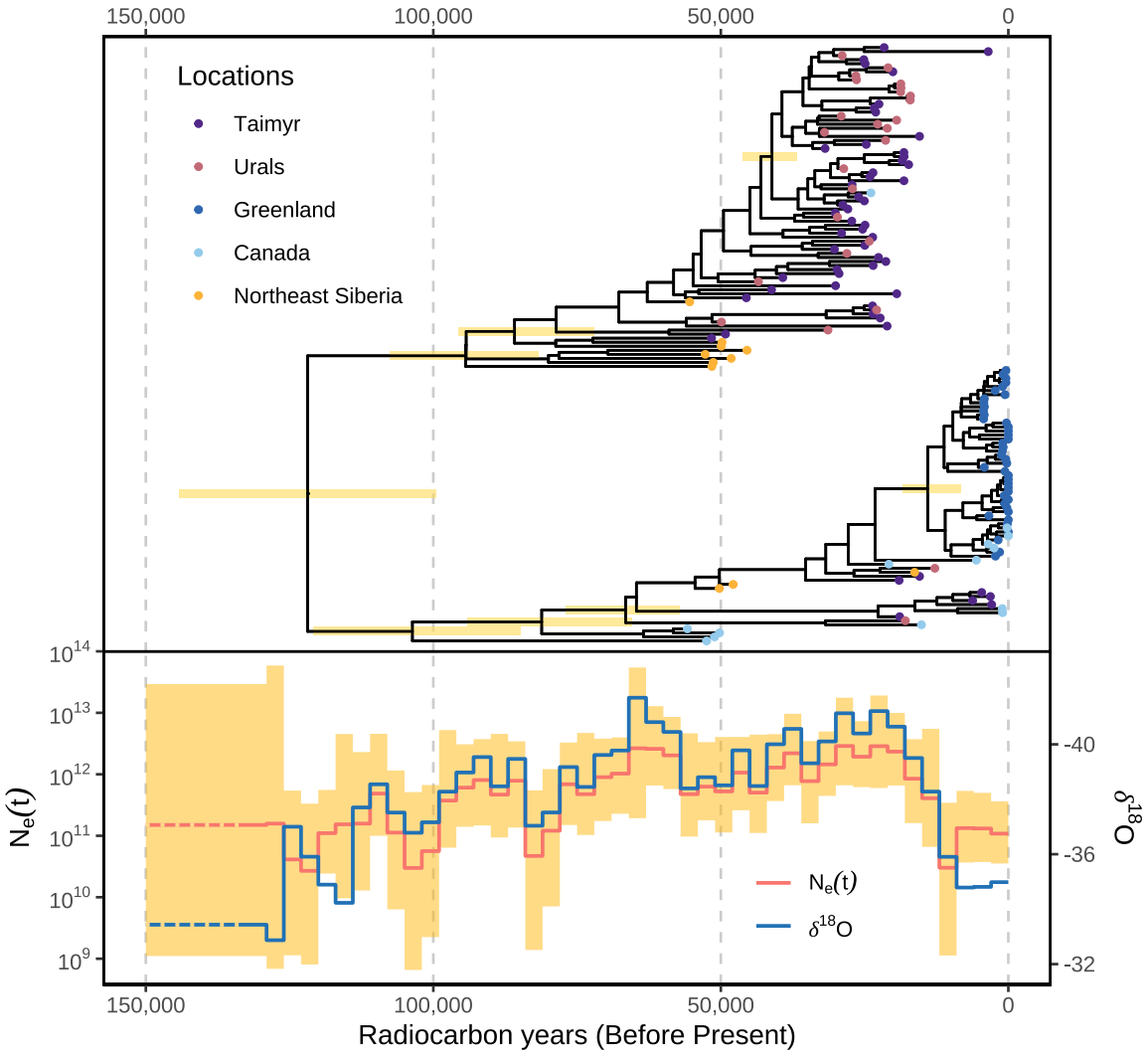

FIGURE 4: ANCIENT MUSK OX: TREE AND DEMOGRAPHY

注：图的上半部分展示了晚第四纪麝牛的进化树，所用标本的年代范围为现今至放射性碳年距今 56,900 年（YBP）。枝端颜色表示样本采集地点，黄色阴影水平矩形表示部分节点年龄的 95%最高后验密度区间。下方折线图描绘了在高斯过程模型下拟合的古代麝牛中位数 N_e（t）水平的趋势（红线），以及格陵兰冰芯计划（Dansgaard 等人，1993；Grootes 等人，1993）的冰芯 $\delta^{18}O$ （蓝线）水平。是氧同位素组成的度量，可作为温度的代用指标（Gill 等人，2016年）。因此，该图表明温度与种群增长之间存在反向关系（请注意，协变量轴是递减的，以突显相关性）。黄色阴影区域表示 N_e（t）的 95%最高后验密度区间。左侧的线条为虚线，因为假定 $\delta^{18}O$ 水平保持恒定，且等于 Gill 等人（2016年）中最早可得的数值。

HIV-1 CRF的一种类型，在全球HIV感染病例中仅占很小比例（7.7%），但却占喀麦隆病例的大多数（60–70%）（Powell等人，2010）。在此，我们利用336份样本的多位点比对，研究该HIV毒株在喀麦隆的种群历史。这些样本是1996至2004年间从雅温得和杜阿拉的献血者中采集的，包含336条gag、pol和env CRF02 AG基因序列（Brennan等人，2008年）。我们的研究以Faria等人(2012)和Gill 等人(2013)的先前研究为基础。前述作者采用基于合并理论的分段常数模型来描述HIV种群大小；此外，我们也借鉴了Gill 等人(2016年)对该模型的扩展，其中推断过程由对数线性模型提供支持，并将喀麦隆成年人（18至49岁）的年度HIV发病率作为协变量。

这个示例之所以特别值得关注，有两个原因：首先，其中包含三个彼此分离的基因，这意味着我们属于第2.3节介绍的多位点框架；其次，我们处于存在缺失协变量值的情形，因为只有1990年之后的发病率可用。对于1990年之前缺失的发病率，我们通过为其指定一个以最早观测值为锚点的高斯马尔可夫随机场先验来处理。然后，我们使用哈密顿蒙特卡罗从缺失协变量的后验分布中进行抽样。

采用更为灵活的发病率信息引导的高斯过程模型，得到了与先前研究一致的结果。具体而言，该HIV毒株的有效群体大小持续增长，直至在1997年达到峰值，此后稳步下降。（图5）。我们的分析还证实了CRF02 AG的种群历史与HIV发病率之间存在密切关联，这一关联见于Gill 等人（2016年），尤其支持了对数线性假设的有效性（图2(c)）。这一关系在图5中也很明显，该图显示推断的种群大小与发病率随时间呈现相似的变化模式。

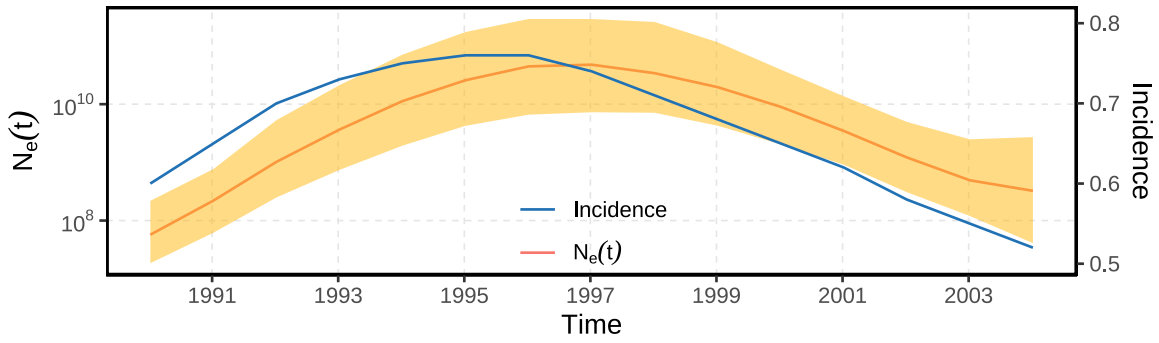

FIGURE 5: HIV EXAMPLE: DEMOGRAPHY

注：折线图展示了在高斯过程模型下学习得到的喀麦隆HIV CRF02 AG毒株有效种群大小水平中位数的趋势-(红线），以及当地总体HIV发病率（蓝线）。这些 $N_e(t)$水平是使用高斯过程（GP）模型推断的，并将发病率作为协变量纳入。推断是在合并先验下假设随机树进行的，其中分段常数 $N_e(t)$使用以发病率为协变量的高斯过程进行建模。黄色阴影区域表示 $N_e(t)$水平的95%最高后验密度区间。图中的视觉趋势表明发病率与 $N_e(t)$水平之间存在直接关系。

4 讨论

我们开发了一个灵活的基于合并过程的推断框架，用高斯过程先验取代对数线性协变量模型，从而能够检测协变量与有效群体大小之间复杂的非线性关系。我们的方法针对现有方法中的一个关键局限：即假设协变量以严格的（对数）线性方式影响人口动态。当这一假设不成立时，线性模型便无法妥善利用协变量中包含的信息来改进整体推断，并会对它们的效应产生有偏估计。

实证见解与方法性能。我们通过三个实证应用展示了本方法，证明了其灵活性带来的优势。在黄热病毒分析中，我们发现温度与病毒种群动态之间呈S形关系，其中两个“趋平点”界定了一个温度区间，在此区间内，升温会推动种群增长。在此范围之外，温度

升高对种群动态的影响微乎其微——而对数线性模型却系统性地误判了这一模式，在不同气温水平下高估或低估其影响。重要的是，我们的研究结果总体上与以下事实相符：蚊媒病毒传播高度依赖温度，并呈单峰模式：传播适宜性随温度升高而增加，直至达到最佳区间，随后由于蚊虫存活率降低和生理胁迫，在更高温度下有所下降（Bel-，2020年；de Souza 和 Weaver，2024年；de Abreu 等人，2025）。

对于古代麝牛，基于高斯过程的模型揭示了温度效应大小的不确定性随温度值而异：在较低温度下，变暖与种群大小之间的反向关系有充分证据支持；但在较高温度下，这种关系存在很大不确定性，而线性模型僵化的随机结构会掩盖这一细微差别。相反，我们对 HIV-1 的分析证实，种群大小与发病率水平之间呈对数线性关系。当真实关系近似线性时，基于高斯过程的模型能够将其还原，表明该模型对其线性对应模型具有良好的泛化能力。

何时使用灵活的协变量模型。我们建议研究人员在以下情况下考虑使用基于高斯过程的模型：(1) 先验知识表明关系可能呈非线性或存在阈值效应（例如物种热耐受范围、疾病传播最适温度）；(2) 初步线性模型拟合不佳或呈现残差模式；(3) 协变量取值范围较广，效应可能随之变化；或 (4) 对协变量不同取值下的不确定性进行量化在科学上十分重要。当协变量缺乏信息，或已明确证明关系为线性时，仍可采用较简单的模型。

局限性与注意事项。模型的有效性在很大程度上取决于协变量的质量和相关性。当 $N_e(t)$在短时间内与某个协变量高度相关时，可能会出现潜在问题，这一点已由 Gill 等人(2016年)指出。

事实上，这可能导致对其他时期协变量效应的外推效果不佳，从而造成错误的人口统计重建。Gill 等人(2016年已经指出，除非其他时期的序列数据信息较弱，否则这种影响有限。我们的高斯过程框架还能通过适应协变量范围内不断变化的不确定性量化，进一步缓解这一问题。

未来方向。本文为多个新的研究方向创造了条件。在方法学方面，尽管我们提出的示例每次只涉及一个协变量，但该框架可自然扩展至多个协变量，从而推动自动化协变量选择程序的发展。通过层次建模将协变量中的测量误差纳入分析，也能进一步改进推断。我们处理缺失协变量的方法（如 HIV 分析中所采用的、以观测值为锚点的高斯马尔可夫随机场先验）可加以形式化，并为此目的进一步拓展。在科学研究方面，该方法有助于开展旨在检测人口变异复杂驱动因素的研究，例如环境因素。

更广泛的影响。通过高斯马尔可夫随机场和高斯过程先验统一时间信息与协变量信息，我们的框架能够更准确、更可解释地重建种群历史。检测非线性协变量效应、量化不确定性如何随协变量范围变化以及处理缺失数据的能力，使该方法对于理解种群如何响应环境波动等外源性驱动因素尤为重要——这一问题在系统发育动力学、疾病生态学和保护中具有核心意义。

致谢

我们衷心感谢超威半导体公司捐赠用于本研究的并行计算资源所提供的支持。

资金支持

本研究得到美国国立卫生研究院资助，资助项目编号为 U19 AI135995、R01 AI153 044, R01 AI162611，以及 R01 AI192139。PL 感谢弗兰德斯研究基金会（弗兰德斯科学研究基金会；G0D5117N、G0B9317N 和 G051322N）的支持。MUGK 感谢以下资助：洛克菲勒基金会（PC-2022-POP-005）、Google.org 健康人工智能计划、牛津马丁学院大流行基因组学与数字化大流行防备计划、欧盟地平线欧洲计划下的 MOOD（编号 874850）和 E4Warning（编号 101086640）项目、惠康信托基金资助项目 303666/Z/23/Z、226052/Z/22/Z 和 228186/Z/23/Z、英国研究与创新署（编号 APP8583）、医学研究基金会（MRF-RG-ICCH-2022-100069）、英国国际发展部（编号 301542-403）、比尔及梅琳达·盖茨基金会（INV063472 和 INV-090281）以及诺和诺德基金会（NNF24OC0094346）的资助。NRF 感谢英国医学研究理事会—圣保罗研究基金会（FAPESP）CADDE合作项目奖（MR/S0195/1 和 FAPESP 18/14389-0）、惠康信托基金气候敏感型传染病建模数字技术开发奖（226075/Z/22/Z）、惠康信托基金登革热与寨卡网络计划（登革热与寨卡免疫学和基因组学多国网络；316633/Z/24/Z）、淡马锡基金会（UNITEDengue）以及英国医学研究理事会全球传染病分析中心（MR/R015600/1）的支持；该中心由英国医学研究理事会（MRC）和英国外交、联邦及发展事务部（FCDO）根据 MRC/FCDO

协约共同资助，并作为由欧盟支持的 EDCTP2计划的一部分。本出版物的内容完全由作者负责，并不一定反映欧盟委员会或其他资助方的观点。

数据可用性

本研究的所有数据、代码和分析文件均可从 GitHub 获取（https://github.com/suchard-group/NonparametricCoalescentProcesses）。该仓库包含序列比对、协变量数据、BEAST X 配置文件、分析脚本，以及本研究中所有示例的结果。

利益冲突

作者声明不存在利益冲突。

参考文献

Ayres, D. L.、Cummings, M. P.、Baele, G.、Darling, A. E.、Lewis, P. O.、Swofford, D. L.、Huelsenbeck, J. P.、Lemey, P.、Rambaut, A. 和 Suchard, M. A.（2019年）。BEAGLE 3：面向统计系统发育学的高性能计算库在性能、扩展性与易用性方面的改进。《系统生物学》，68(6):1052–1061。

Baele, G.、Gill, M. S.、Lemey, P. 和 Suchard, M. A.（2020年）。在贝叶斯系统发育学框架下，使用skygrid溯祖模型进行哈密顿蒙特卡罗抽样以估计过去种群动态。*Wellcome Open Research*，5:53。

Baele, G., Ji, X., Hassler, G. W., McCrone, J. T., Shao, Y., Zhang, Z., Holbrook, A. J.,

Lemey, P., Drummond, A. J., Rambaut, A., and Suchard, M. A. (2025). BEAST X 用于贝叶斯系统发育、系统地理学和系统动力学推断。《自然·方法》，第 1–4 页。出版方：自然出版集团。

Bellone, R. 和 Failloux, A.-B.（2020年）。温度在塑造蚊媒病毒传播中的作用。*Front Microbiol*，11:584846。

Beskos, A.、Pillai, N.、Roberts, G.、Sanz-Serna, J.-M. 和 Stuart, A.（2013）。混合蒙特卡洛算法的最优调节。*Bernoulli*，19(5A):1501–1534。出版商：伯努利数理统计与概率学会。

Brennan, C. A.、Bodelle, P.、Coffey, R.、Devare, S. G.、Golden, A.、Hackett, J. J.、Harris, B.、Holzmayer, V.、Luk, K.-C.、Schochetman, G.、Swanson, P.、Yamaguchi, J.、Vallari, A.、Ndembi, N.、Ngansop, C.、Makamche, F.、Mbanya, D.、G¨urtler, L. G.、Zekeng, L. 和 Kaptu´e, L.（2008年）。1996年至2004年喀麦隆献血者中多种HIV-1毒株的流行率保持稳定。《*JAIDS*：获得性免疫缺陷综合征杂志》，49(4):432。

Campos, P. F., Willerslev, E., Sher, A., Orlando, L., Axelsson, E., Tikhonov, A., Aaris- Sørensen, K., Greenwood, A. D., Kahlke, R.-D., Kosintsev, P., Krakhmalnaya, T., Kuznetsova, T., Lemey, P., MacPhee, R., Norris, C. A., Shepherd, K., Suchard, M. A., Zazula, G. D., Shapiro, B., and Gilbert, M. T. P.（2010）。古DNA分析排除了人类是晚更新世麝牛（Ovibos moschatus）种群动态背后驱动力的可能性。《美国国家科学院院刊》，107(12):5675–5680。出版方：《美国国家科学院院刊》。

Dansgaard、Johnsen, S. J.、Clausen, H. B.、Dahl-Jensen, D.、Gundestrup, N. S.、Hammer, C. U.、Hvidberg, C. S.、Steffensen, J. P.、Sveinbj¨ornsdottir, A. E.、Jouzel, J. 和 Bond, G.（1993）。来自一份

25万年冰芯记录的过去气候普遍不稳定性的证据。《自然》，364(6434):218–220。出版方：自然出版集团。

de Abreu, R., Fernandez, I. P., Mishra, S., Gutierrez, B., Inward, R. P. D., Mills, C., Ortiz, E. L., Bastos, L. S., Freitas, L. P., Carvalho, L. M., Flaxman, S., Bhatt, S., Scarpino, S. V., Coelho, F. C., Reiner, R.C., Sambaturu, P., Tegally, H., Cauchemez, S., Cruz, O. G., Munayco, C. V., D´ıaz-Qui˜nonez, J. A., Mitchell, D., Lott, F., Dominici, F., Pybus, O. G., Code¸co, C. T., Castro, M. C., Kraemer, M. U. G., and Sparrow, S. (2025). 气候变化在登革热扩张中的作用。$medRxiv$。

de Souza, W. M. 和 Weaver, S. C.（2024年）。气候变化和人类活动对媒介传播疾病的影响。《自然评论：微生物学》，22(8)：476–491。

Drummond, A. J.、Nicholls, G. K.、Rodrigo, A. G. 和 Solomon, W.（2002年）。利用按时间间隔采集的序列数据，同时估计突变参数、种群历史和谱系。《遗传学》，161(3)：1307–1320。

Drummond, A. J.、Rambaut, A.、Shapiro, B. 和 Pybus, O. G.（2005）。基于分子序列的过去种群动态贝叶斯合并推断。《分子生物学与进化》，22(5)：1185–1192。

Faria, N. R.、Suchard, M. A.、Abecasis, A.、Sousa, J. D.、Ndembi, N.、Camacho, R.、Vandamme, A.-M.、Peeters, M. 和 Lemey, P.（2012）。喀麦隆 HIV-1 CRF02 AG 分支的系统发育动力学。《感染、遗传与进化：传染病分子流行病学与进化遗传学期刊》，12(2)：453–460。

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., 和 Rubin, D. B.（2013）。《贝叶斯数据分析》。Chapman and Hall/CRC，纽约，第3版。

Gill, M. S., Lemey, P., Bennett, S. N., Biek, R., 和 Suchard, M. A.（2016年）。理解过去种群动态：基于协变量的基于溯祖的贝叶斯建模。系统生物学，65(6)：1041–1056。

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B. 和 Suchard, M. A. (2013). 改进贝叶斯种群动态推断：一种适用于多位点的基于溯祖的模型。《分子生物学与进化》，30(3):713–724.

Girolami, M. 和 Calderhead, B.(2011). 黎曼流形朗之万与哈密顿蒙特卡洛方法。英国皇家统计学会会刊*B*辑：统计方法学，73(2):123–214.

Green, P. J. (1995). 可逆跳跃马尔可夫链蒙特卡洛计算与贝叶斯模型确定。*Biometrika*，82(4):711–732. 出版商：[牛津大学出版社，Biometrika Trust]。

Griffiths, R. C. 和 Tavar´e, S. (1994). 变化环境中的中性等位基因抽样理论。哲学汇刊：生物科学，344(1310):403–410. 出版商：英国皇家学会。

Grootes, P. M.、Stuiver, M.、White, J. W. C.、Johnsen, S. 与 Jouzel, J.（1993）。GISP2 与 GRIP 格陵兰冰芯氧同位素记录的比较。《自然》，366(6455):552–554。出版者：自然出版集团。

Hill, S. C., Dellicour, S., Claro, I. M., Sequeira, P. C., Adelino, T., Th´ez´e, J., Wu, C.-H., Moreira, F. R. R., Giovanetti, M., Li, S. L., Jesus, J. G. d., Col´on-Gonz´alez, F. J., Chamberlain, H. R., Pannell, O., Tejedor-Garavito, N., Bruycker-Nogueira, F. d., Fabri, A. A., Mares-Guia, M. A., Xavier, J., Zarebski, A. E., Hamlet, A., Sallum, M.A.M., Costa, A.C. d., Manuli, E. R., Levin, A. S., Mucci, L.F., Tubaki, R. M., Menezes, R. M. T. d., Deus, J. T. d., Spinola, R., Saad, L., Kallas, E. G.,

Wint, G. R. W., Peixoto, P. S., Santos, A. A. d. S., Messina, J. P., Brady, O. J., Tatem, A. J., Suchard, M. A., Mendez-Rico, J. A., Abreu, A., Aguiar, R. S., Pybus, O. G., Baele, G., Lemey, P., Iani, F., Cunha, M. S., Filippis, A. M. B. d., Sabino, E. C., and Faria, N. R. (2022年). 气候与土地利用塑造了人畜共患黄热病病毒的传播。*medRxiv*。

Hudson, R. R.（1990年）。基因谱系与合并过程。《牛津进化生物学评论》，7(1):44。

Jukes, T. H. 和 Cantor, C. R.（1969年）。蛋白质分子的进化。载于《哺乳动物蛋白质代谢》，第21–132页。爱思唯尔。

Kimura, M.（1980年）。通过核苷酸序列的比较研究估算碱基替换进化速率的一种简单方法。《分子进化杂志》，16(2):111–120。

Kingman, J. F. C.（1982年）。共祖过程。《随机过程及其应用》，13(3):235–248。

Kuhner, M. K.、Yamato, J. 和 Felsenstein, J.（1998年）。基于共祖过程的种群增长率最大似然估计。《遗传学》，149(1):429–434。

Lorenzen, E. D., Nogu´es-Bravo, D., Orlando, L., Weinstock, J., Binladen, J., Marske, K. A., Ugan, A., Borregaard, M. K., Gilbert, M. T. P., Nielsen, R., Ho, S. Y. W., Goebel, T., Graf, K. E., Byers, D., Stenderup, J. T., Rasmussen, M., Campos, P. F., Leonard, J. A., Koepfli, K.-P., Froese, D., Zazula, G., Stafford, T. W., Aaris- Sørensen, K., Batra, P., Haywood, A. M., Singarayer, J. S., Valdes, P. J., Boeskorov, G., Burns, J. A., Davydov, S. P., Haile, J., Jenkins, D. L., Kosintsev, P., Kuznetsova, T., Lai, X., Martin, L. D., McDonald, H. G., Mol, D., Meldgaard, M., Munch, K., Stephan, E., Sablin, M., Sommer, R. S., Sipko, T., Scott, E., Suchard, M. A.,

Tikhonov, A., Willerslev, R., Wayne, R. K., Cooper, A., Hofreiter, M., Sher, A., Shapiro, B., Rahbek, C., and Willerslev, E.（2011）。晚第四纪巨型动物群对气候和人类的物种特异性响应。《自然》，479(7373)：359–364。出版商：自然出版集团。

Metropolis, N.、Rosenbluth, A. W.、Rosenbluth, M. N.、Teller, A. H. 和 Teller, E.（1953）。利用快速计算机计算状态方程。《化学物理杂志》，21(6):1087–1092。

Minin, V. N.、Bloomquist, E. W. 和 Suchard, M. A.（2008年）。穿越崎岖 Skyline 的平滑 Skyride：基于溯祖的种群动态贝叶斯推断。《分子生物学与进化》，25(7):1459–1471。

Monti, F.、Ji, X. 和 Suchard, M. A.（2025）。连续时间马尔可夫链的非参数建模。*arXiv* 预印本 *arXiv:2511.03954*。

Neal, R. M.（2011）。使用哈密顿动力学进行马尔可夫链蒙特卡洛。载于《马尔可夫链蒙特卡洛手册》，第113–162页。Chapman and Hall/CRC，纽约，第1版。

Opgen-Rhein, R.、Fahrmeir, L. 和 Strimmer, K.（2005）。使用可逆跳跃马尔可夫链蒙特卡洛从谱系树推断人口史。《*BMC*进化生物学》，5(1):6。

Palacios, J. A. 和 Minin, V. N.（2013）。基于高斯过程的贝叶斯非参数方法：从基因谱系推断种群大小轨迹。《生物统计学》，69(1):8–18。

Powell, R., Barengolts, D., Mayr, L. 和 Nyambi, P.（2010）。《喀麦隆农村地区 HIV- 1 多样性的演化及其对疫苗设计和试验的影响》。《病毒》，2(2):639–654。出版商：分子多样性保护国际。

Pybus, O. G.、Rambaut, A. 和 Harvey, P. H.（2000）。《从重建的谱系推断病毒种群历史的整合框架》。《遗传学》，155(3):1429–1437。

Rasmussen, C. E. 和 Williams, C. K. I. (2006). 用于机器学习的高斯过程. 自适应计算与机器学习. MIT出版社，剑桥，马萨诸塞州。OCLC: ocm61285753.

Strimmer, K. 和 Pybus, O. G. (2001年). 使用广义天际线图探索DNA序列的人口史。《分子生物学与进化》，18(12):2298–2305.