arXiv:2602.06135v1 [stat.AP] 5 Feb 2026

使用具有跨辖区滞后的套索向量自回归模型对美国辖区猴痘疫情暴发进行早期预警

2026年2月9日

Hannah Craddock[1],**Joel O. Wertheim**[1],**Eliah Aronoff-Spencer**[1],**Mark Beatty**[2],**David Valentine**[1],**Rishi Graham**[1],**Jade C. Wang**[3],**Lior Rennert**[4],**Seema Shah**[2*],**Ravi Goyal**[1*],**Natasha K. Martin**[1*]

[1] 加利福尼亚大学圣迭戈分校传染病与全球公共卫生部，美国加利福尼亚州圣迭戈[2] 圣迭戈县卫生与公共服务局，美国加利福尼亚州圣迭戈[3] 加利福尼亚大学圣迭戈分校生物信息学与系统生物学研究生项目，美国加利福尼亚州拉霍亚[4] 克莱姆森大学行为、社会与健康科学学院公共卫生系，美国南卡罗来纳州克莱姆森

* 共同资深作者

摘要

猴痘是一种正痘病毒属病毒，可感染人类和动物，主要通过密切的身体接触传播。猴痘传播呈现间歇性且空间分布不均的动态特征，这凸显了及时提供特定地区预测的必要性，以支持美国有针对性的公共卫生响应。我们使用美国疾病控制与预防中心（CDC）的全国监测数据，开发了一种采用 Lasso 正则化的向量自回归模型（VAR-Lasso），为美国八个高发病率辖区生成滚动的提前两周每周猴痘病例预测。VAR-Lasso模型识别出显著的长滞后期跨辖区预测变量。以圣迭戈县（SDC）为案例研究，这些统计预测变量与系统发育分析结果相符；系统发育分析追溯发现，SDC在2023年的一个簇源自六个月前伊利诺伊州的一次疫情暴发。由于发病率上升时往往最需要采取公共卫生行动，我们的性能评估重点关注正斜率加权误差指标。我们将VAR-Lasso模型的预测性能与单变量自回归（AR）Lasso模型以及朴素移动平均估计进行比较。比较指标包括斜率加权均方根误差（RMSE）、斜率加权平均绝对误差（MAE）和斜率加权偏差。在所有观测中，与AR模型相比，VAR-Lasso模型的斜率加权RMSE、MAE和偏差分别降低了12%、7%和66%；与朴素基准相比，则分别降低了16%、13%和76%。我们的研究结果突显了稀疏多变量时间序列模型的价值：这类模型利用跨辖区病例数据，能够及早预测猴痘疫情暴发。此类预测有助于卫生部门主动及时地提供资源并发布信息，以降低未来疫情暴发的风险。

1 引言

猴痘是一种人畜共患传染病，由猴痘病毒（MPXV）引起。猴痘病毒是一种正痘病毒属双链 DNA 病毒，分为两个支系（I支和II支）（Mitj`a et al., 2023）。2022年，猴痘IIb支在全球大规模暴发，并持续传播至今。2022年的疫情影响了此前报告病例很少或从未报告病例的地区，包括美国和欧洲。在此次疫情中，男男性行为者（MSM）的猴痘发病率明显高于其他人群（Laurenson-Schafer et al., 2023）。这与撒哈拉以南非洲地区观察到的典型流行病学特征不同，表明在高收入国家，猴痘可能主要在人际身体接触频繁的人群中传播，尤其是男男性行为者（Brand et al., 2023）。针对猴痘的多方面应对举措包括扩大疫苗接种的可及性，并制定面向男男性行为者的定向宣传信息，以降低其感染风险（Delaney, 2022）。目前已有一种效果显著的疫苗 Jynneos（Pischel et al., 2024），多个辖区在2022年疫情暴发初期启动了疫苗接种运动（Owens, 2023）。尽管美国的大规模疫情在2022年末趋于平息，但多个辖区仍持续出现较小规模的疫情暴发，以及发病率间歇性升高的时期（美国疾病控制与预防中心, 2025）。近期，2025年10月，加利福尼亚州报告了美国首批与社区传播相关的I支病例（加利福尼亚州公共卫生部, 2025；美国疾病控制与预防中心, 2025）。这些进展凸显了持续建设强有力的早期预警系统的必要性，以支持及时发现疫情并采取公共卫生应对措施。

预测模型已成为疫情暴发检测与应对的关键工具。从公共卫生角度来看，能够提供早期预警信号并提前预判病例增加十分重要。此类早期预警预测可帮助公共卫生部门有效规划，包括协调疫苗接种策略和健康教育宣传活动（Overton等人，2023年）。单变量自回归（AR）时间序列模型是预测猴痘的常用方法（Munir等人，2024年；Bakare等人，2025年）。然而，这类模型无法考虑多个辖区之间相互依赖的趋势。向量自回归（VAR）模型将AR方法扩展至多变量情境（Stock和Watson，2001年），能够对动态的跨辖区关系进行建模。然而，随着辖区数量和滞后项增加，VAR模型的参数会变得非常稠密。最小绝对收缩与选择算子（Lasso）等正则化方法通过将部分系数收缩至恰好为零来应对这一挑战，从而有效地进行变量选择（Tibshirani，1996年）。因此，VAR-Lasso框架既能进行预测，也能识别具有影响力的跨辖区滞后预测变量。

据我们所知，尽管 VAR-Lasso 具有潜力，但此前尚无研究将其用于猴痘预测。本研究使用美国疾病控制与预防中心（CDC）的数据，应用 VAR-Lasso模型预测美国八个高发病率辖区的每周猴痘病例数。我们以加利福尼亚州圣迭戈县（SDC）为例开展了详细的案例研究；圣迭戈县是这八个高发病率辖区之一，本研究也得到了与圣迭戈县卫生与公共服务局（CSDHHSA）持续合作的支持。这项合作是美国疾病控制与预防中心旗下 Insight Net 的一部分。Insight Net 是一个由美国各地 13 个疫情分析与建模中心组成的网络，Resilient Shield（加利福尼亚大学圣迭戈分校）

与 CSDHHSA 合作。本文介绍的工作是开发早期预警系统的一部分，旨在检测猴痘发病率的上升；因此，SDC 为评估和验证预测模型提供了关键的实际应用背景。

在预测评估中，某些流行病学特征可能比其他特征更为重要，因此在评估预测性能时，可以相应地对其赋予不同权重（Tabataba 等人，2017；Armstrong，2001）。由于发病率上升期间往往最需要采取公共卫生行动，我们提出斜率加权误差指标，根据逐周正增长的幅度对预测误差加权，确保在模型评估中，传播增加的时段得到相应更高的重视。

本研究的主要目标是利用报告病例数据中跨辖区的关联，提高美国高发病率辖区在病例增加期间进行猴痘病例预测的及时性和准确性。为此，我们：（1）开发一种包含跨辖区滞后项的 VAR-Lasso 预测模型，对美国八个高发病率辖区的每周猴痘病例数进行提前两周的预测；（2）通过开展系统发育分析，将模型识别出的滞后预测变量与 SDC 的系统发育发现进行比较，从而对模型进行外部验证；（3）在病例增加期间，使用反映发病率上升时公共卫生行动需求加剧的正斜率加权误差指标评估模型性能；（4）将 VAR-Lasso 预测与 AR-Lasso 模型及朴素移动平均估计进行基准比较。综合而言，这些步骤开发、验证并评估了一个由系统发育证据支持的稀疏多元时间序列框架。该框架整合跨辖区病例信息并捕捉长滞后传播信号，从而提升短期猴痘预测能力。

2 结果

首先，我们提供美国猴痘病例的总体描述性统计，并重点介绍八个高发病率辖区。其次，我们评估 VAR-Lasso 模型对 2024 年所有预测的整体预测表现，并将结果与单变量 AR Lasso 模型和朴素移动平均估计器进行比较。第三，我们对 SDC 开展案例研究，详细评估其预测，并使用独立的系统发育数据验证模型预测。第四，我们报告其余辖区的预测结果，并总结模型表现。最后，我们提供其他敏感性分析。

2.1 描述性统计——美国猴痘病例与高发病率辖区

2022年猴痘大规模全球暴发，导致美国报告了30,069例病例，遍及全部50个州。此后，2023年报告了1,747例病例，2024年1月至11月报告了2,529例病例（美国疾病控制与预防中心公布的最新数据）。我们将建模重点放在发病率最高的八个辖区（包括州、县和市）上；如表 1所示，这些辖区在2022年至2024年间的病例数占全部病例的55%。这八个辖区分别是纽约市、得克萨斯州、洛杉矶县、佛罗里达州、伊利诺伊州、佐治亚州、圣迭戈县和华盛顿州。我们重点建模的时间段为2023年1月至2024年11月。每周报告的病例数

辖区	年度病例总数（占美国年度总数的百分比）		
	2022	2023	2024年*
纽约市	3837（13%）	198（11%）	392（16%）
得克萨斯州	2969（10%）	187（10%）	283（11%）
洛杉矶县	2281 (8%)	120 (7%)	202 (8%)
佛罗里达州	2858 (10%)	109 (6%)	178（7%）
伊利诺伊州	1424（5%）	145（8%）	77（3%）
佐治亚州	1986（7%）	90（5%）	70（3%）
圣迭戈县	471（1.5%）	59（3%）	77（3%）
华盛顿州	653（2%）	81（5%）	47（2%）
排名前八的病例总数	**16,479（55%）**	**989（57%）**	**1326（52%）**
美国病例总数	**30,069**	**1,747**	**2,529**

表1：2022–2024年各辖区猴痘病例数及其占美国病例总数的比例（选定的高发病率辖区）。辖区按2023–2024年（模型所涵盖的时期）的病例总数排序。*请注意，2024年的汇总数据涵盖2024年1月至 17th 2024 开始的那一周（美国疾病控制与预防中心提供的最新数据）。

在此期间，发病率最高的八个辖区的病例数见图 1。

2.2 VAR-Lasso预测性能与基准模型比较

我们评估了 VAR-Lasso 模型对 2024 年美国八个高发病率辖区每周猴痘病例数的预测性能，并将其与单变量 AR-Lasso 模型和朴素¨移动平均估计器进行比较（见方法）。模型最初使用 2023年1月至12月的数据进行训练，并根据 2024 年 42 次滚动的提前两周预测评估其样本外表现。随着 2024 年数据逐周增加，训练集也随之更新。在所有预测中，VAR-Lasso 在所有斜率加权指标上始终优于两个基准模型，表明它在捕捉发病率上升方面具有更高的准确性（图2；表2）。VAR-Lasso 模型的斜率加权均方根误差为 1.75，而 AR-Lasso 和朴素¨基准模型分别为 2.0 和 2.11，相对提升分别为 12% 和 16%。
VAR-Lasso 模型的斜率加权均方根误差为 1.75，这表明，以观察到的病例每周正向变化加权后，该模型在各辖区的预测与观察趋势之间的偏差每周不到两例；与 AR-Lasso 和朴素¨基准模型相比，它能更准确地捕捉发病率变化的幅度。斜率加权平均绝对误差为 1.4，而 AR-Lasso 和朴素¨模型分别为 1.6 和 1.5（提升 7% 和 13%）。VAR-Lasso 模型的斜率加权平均绝对误差为 1.42，这表明，以观察到的病例每周正向变化加权后，预测偏差约为

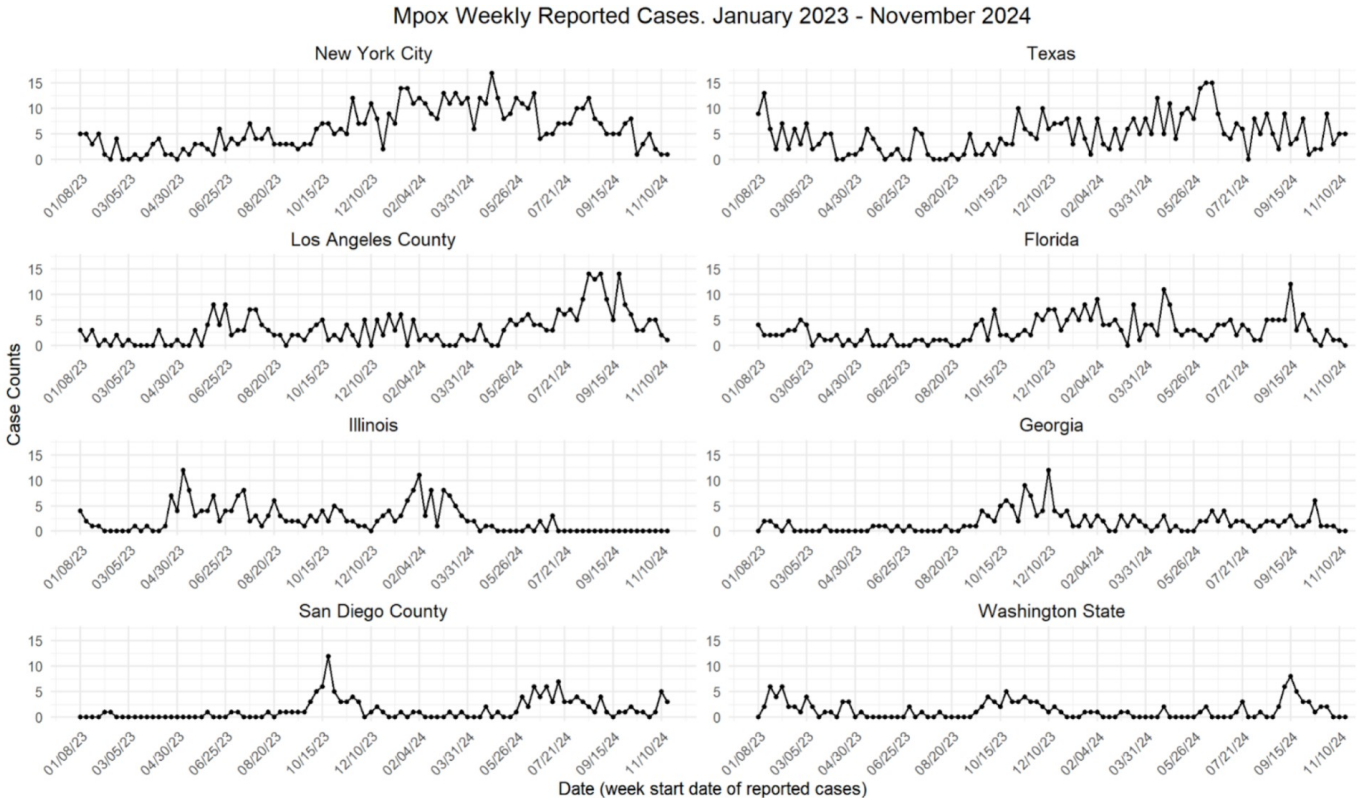

图1：2023年1月至2024年11月美国8个猴痘高发病率辖区每周报告的猴痘病例。数据来源：美国疾病控制与预防中心（CDC）国家监测系统。

模型	评估指标			VAR 的提升百分比	
	VAR-Lasso	AR-Lasso	朴素估计	% 改善$_{VA}$	% 改善$_{VN}$
斜率加权均方根误差	1.75	2.00	2.11	12	16
斜率加权平均绝对误差	1.42	1.52	1.64	7	13
斜率加权偏差	-0.19	-0.56	-0.80	66	76

表2： 2024年1月至11月的42周内，发病率最高的八个辖区所有每周预测的斜率加权误差指标。分别针对使用VAR-Lasso模型、AR-Lasso模型和朴素移动平均估计器生成的预测，计算斜率加权均方根误差、斜率加权平均绝对误差和斜率加权偏差。最后两列显示VAR-Lasso模型预测相较于AR-Lasso模型预测的改善百分比（% 改善$_{VA}$），以及VAR-Lasso模型预测相较于朴素估计预测的改善百分比（% 改善$_{VN}$）。

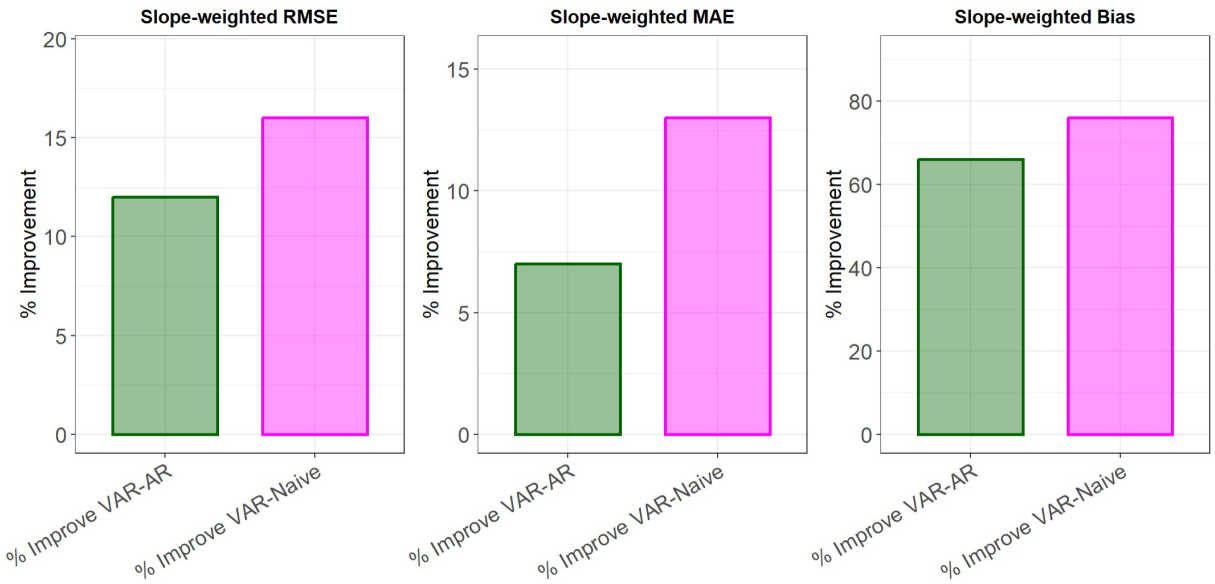

图2：VAR-Lasso模型预测相较于AR-Lasso模型预测及朴¨素移动平均估计预测的百分比改进，基于斜率加权指标。对2024年1月至11月期间42周内，发病率最高的八个辖区的所有每周预测计算斜率加权均方根误差、平均绝对误差和偏差。图中展示了VAR-Lasso模型预测相对于AR-Lasso预测的百分比改进（% Improve VAR-AR），以及相对于朴¨素预测的百分比改进（% Improve VAR-Naive）。

每周1.4例，能够可靠地反映每周发病率变化的幅度。VAR-Lasso模型的斜率加权偏差降至–0.19，而AR-Lasso模型和朴素基准模型分别为–0.56和–0.80（改善66%和76%），表明与另外两个模型相比，其系统性低估程度显著降低。

2.3 案例研究：圣迭戈县——预测结果与系统发育验证

SDC为评估和验证预测模型提供了实际应用场景，这得益于与CSDHHSA的持续合作。2023–2024年，SDC经历了低水平但反复出现的猴痘传播，并间歇性出现发病率升高的时期。其中包括2023年10月某一周每周病例数达到12例的峰值，以及2024年夏季病例数上升、于当年7月达到每周7例的最高值。这些波动使SDC成为评估模型预测病例短期增长能力的理想场景。与总体结果一致，VAR-Lasso模型在SDC的表现优于两种基准模型。与AR模型相比，VAR-Lasso模型的斜率加权均方根误差、平均绝对误差和偏差分别降低了11%、5%和88%；与朴¨素基准模型相比，分别降低了20%、17%和90%（见表3）。

使用 VAR-Lasso、AR-Lasso 和朴素¨估计量得到的 SDC 在 2024年1月至11月期间的每周预测如图 3所示。2024年6月病例数开始上升，并于2024年7月达到每周7例的峰值时，AR-Lasso 和朴素¨估计量均低估了这波激增。相比之下，VAR-Lasso模型能够利用跨辖区的滞后病例数据预测这些峰值，生成与观测到的峰值更为吻合的预测。模型识别出伊利诺伊州是 SDC 最具影响力的外部预测因子，估计系数最大的对应于23周和24周滞后项（图 4a）。23至24周滞后项占主导地位，表明存在长期的跨辖区关联：伊利诺伊州的病例增加大约早于 SDC 的发病率六个月。将伊利诺伊州和 SDC 报告的每周病例数叠加后，可以从原始病例数中清楚地看到这六个月的滞后（图 4b）。具体而言，伊利诺伊州在2023年4月和2024年2月出现的峰值，分别比 SDC 在2023年10月和2024年7月出现的峰值早约23周。

使用系统发育分析进行外部验证。为验证 VAR-Lasso 选出的跨辖区和时间滞后预测变量，我们进行系统发生地理学分析，以推断猴痘病毒传入 SDC 的时间和来源。根据与 SDC 病例相关的 29 个猴痘病毒基因组，我们识别出 2023–2024年期间的五次独立传入事件（表 S1）。最大的簇（B簇）包含 29 个病毒基因组中的 19 个。B簇的 SDC 序列采样于 2023年10月和11月，其最近共同祖先的时间为 2023年9月（图 5）。系统地理学推断表明，猴痘病毒可能来自伊利诺伊州（后验概率 $= 0.72$）、加利福尼亚州（后验概率 $= 0.27$）或洛杉矶县（后验概率 $= 0.02$）。无论该传入病毒的直接祖先所在地为何，该簇的最终来源都是 2023年4月在伊利诺伊州传播的猴痘病毒（后验概率 $= 1.0$）。这表明从伊利诺伊州到 SDC 的时间滞后约为 22–28 周。

这项独立的基因组证据与 VAR-Lasso 的结果高度吻合，VAR-Lasso 将最大效应归于 23–24 周滞后项（图 4a）。统计分析与系统发育分析共同指向相同的跨辖区传播路径，即伊利诺伊州的猴痘活动与 SDC 随后在 23–24 周后出现的高峰相关。

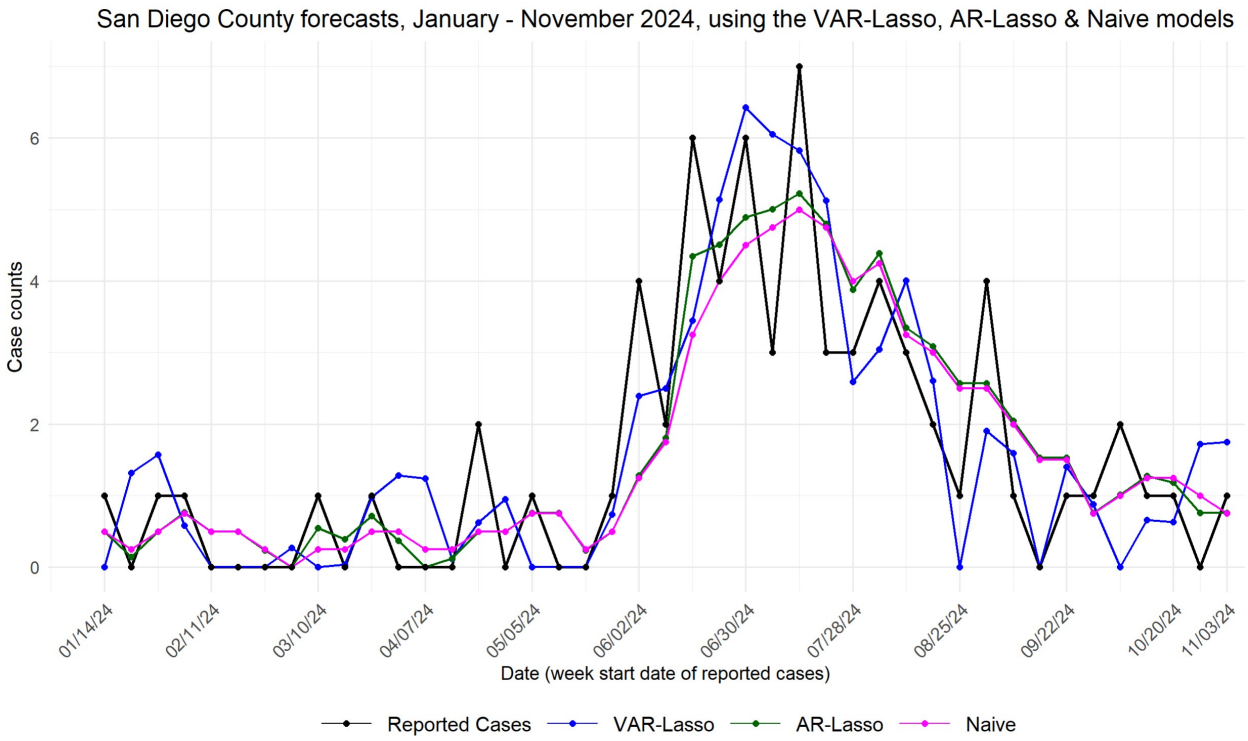

图3：圣迭戈县2024年1月至11月每周报告的猴痘病例及预测。报告病例（黑色）、VAR-Lasso预测（蓝色）、AR-Lasso预测（洋红色）和naïve移动平均预测（绿色）。

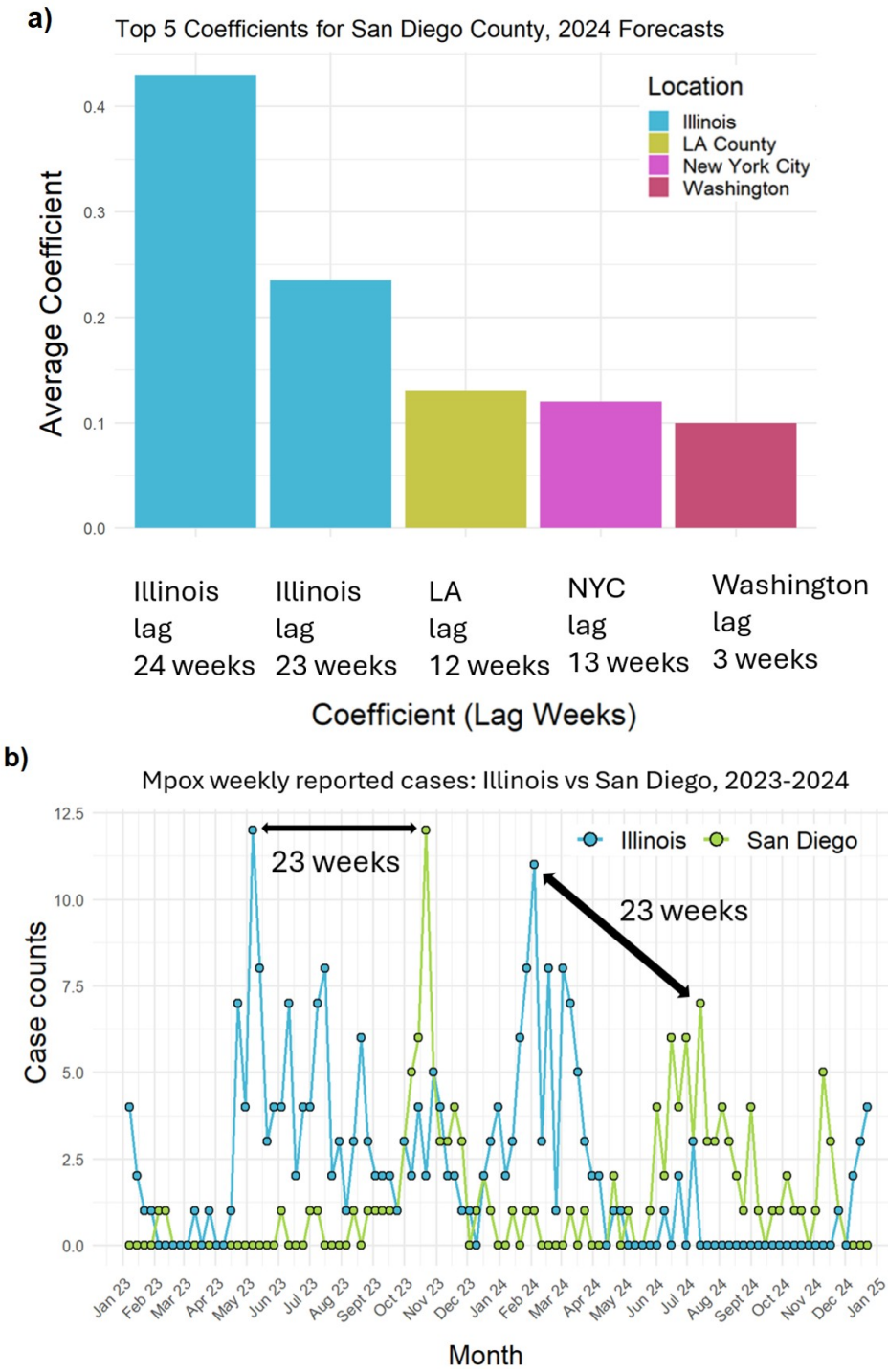

图4：圣迭戈县（SDC）的 VAR Lasso 模型系数前五名，以及 2023—24年 SDC 和伊利诺伊州的每周病例数。（a）。系数表示用于预测 2024年42周病例数的所有 VAR–Lasso 模型拟合结果中的平均系数幅度。每个模型均使用 2023年的全部数据，以及截至预测周前两周可用的数据进行拟合。幅度最大的系数对应伊利诺伊州的 23—24周滞后项，其次是洛杉矶县（12周滞后项）、纽约市（13周滞后项）和华盛顿州（3周滞后项）。（b）。报告的 SDC 和伊利诺伊州每周猴痘病例数（2023–2024年）显示，伊利诺伊州的病例峰值领先 SDC 约 23周，这与来自伊利诺伊州的传入事件的系统发育证据一致。10

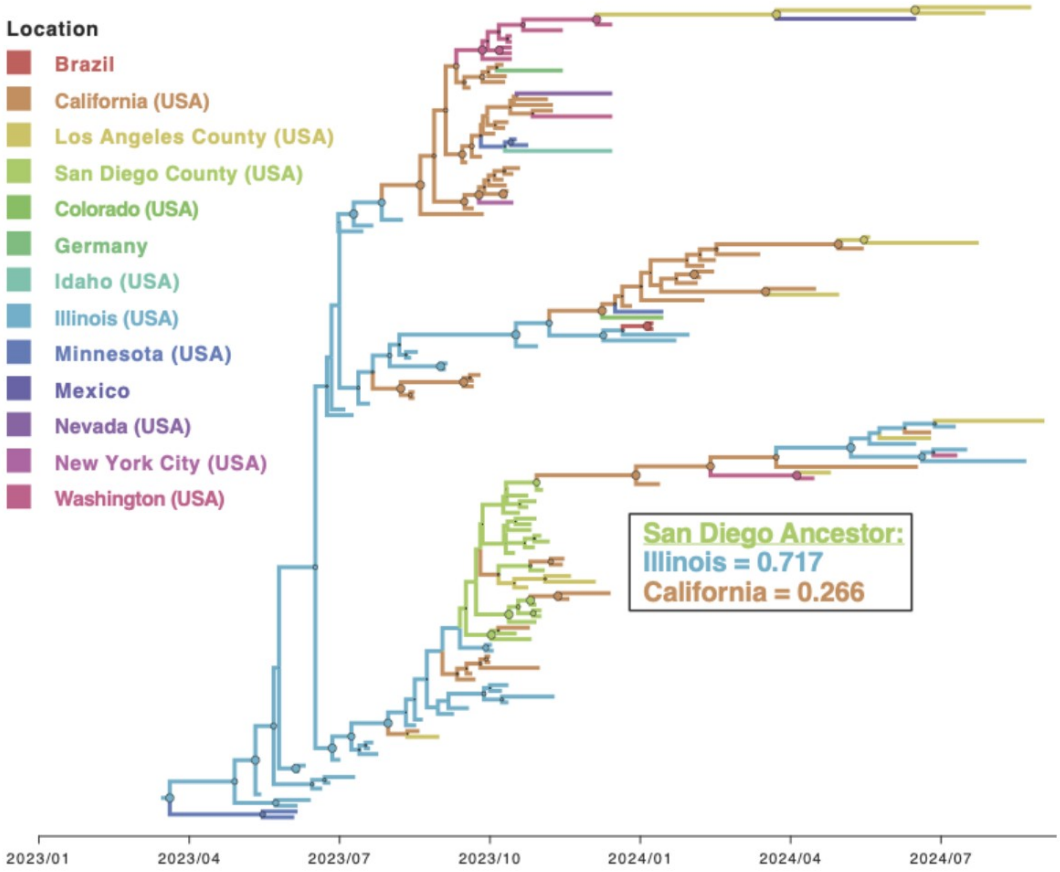

图5：描绘2023年末猴痘病毒迁入圣迭戈县的最大分支可信度树。分支颜色根据推断出的地理位置标示。内部节点圆圈的大小表示后验支持度，颜色表示地理位置。插图标示了圣迭戈县猴痘病毒基因组祖先的直接前序分支所在地的后验概率。

辖区	斜率加权均方根误差					斜率加权平均绝对误差				
	VAR	AR	朴素法	% Imp_{VA}	% Imp_{VN}	向量自回归	自回归	朴素法	% Imp_{VA}	% Imp_{VN}
纽约市	1.50	1.60	2.10	6	26	1.3	1.3	1.7	0	25
得克萨斯州	2.10	2.20	2.30	4	7	1.8	1.8	1.9	0	3
洛杉矶县	1.95	2.60	2.70	25	26	1.7	2.2	2.2	24	24
佛罗里达州	1.80	1.60	1.70	-14	-11	1.5	1.3	1.3	-13	-13
伊利诺伊州	2.10	2.60	2.60	19	18	1.8	2.1	2.0	17	14
佐治亚州	1.10	1.00	1.20	-8	8	0.7	0.9	1.0	13	25
圣迭戈县	1.20	1.35	1.50	11	20	1.0	1.1	1.2	5	17
华盛顿州	1.00	1.80	1.80	44	46	0.7	1.1	1.1	35	40

表3：2024年1月至11月42周内每周预测的辖区级斜率加权误差指标。列中列出了使用VAR-Lasso、AR-Lasso以及na¨ıve移动平均估计器得到的预测的斜率加权均方根误差和平均绝对误差，以及VAR-Lasso模型相较于AR-Lasso模型（% Imp_{VA}）和na¨ıve估计器（% Imp_{VN}）的改进百分比。

2.4 辖区 高发病率辖区的预测性能 s

我们通过计算各辖区在2024年的42次每周预测中的斜率加权误差指标，评估辖区层面的模型表现（图 6）。就斜率加权均方根误差和平均绝对误差而言，VAR-Lasso模型在8个辖区中的6个优于AR-Lasso模型，在8个辖区中的7个优于朴素基准¨（表3）。纽约市、得克萨斯州、洛杉矶县、伊利诺伊州、圣迭戈县和华盛顿州均观察到性能提升。出现提升时，相较于AR-Lasso模型，斜率加权均方根误差降低了4%至44%；相较于朴素基准¨，则降低了7%至46%。斜率加权平均绝对误差的降幅趋势相似：相较于AR-Lasso模型降低了5%至35%，相较于朴素基准¨则降低了3%至40%。只有佛罗里达州的均方根误差和平均绝对误差均出现负向改善。总体而言，这些结果表明，VAR-Lasso模型在辖区层面具有优势。

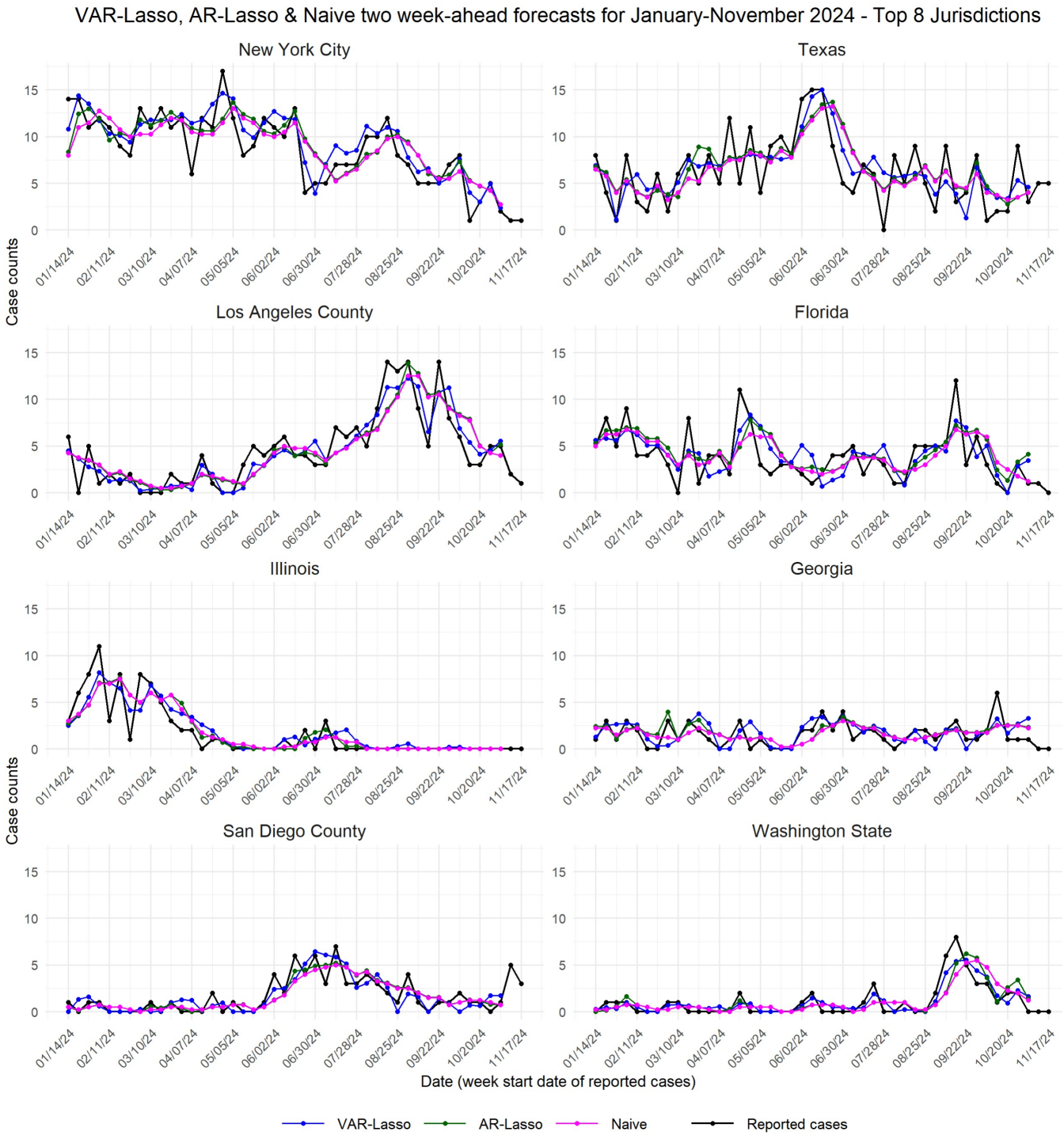

图6：2024年1月至11月美国八个高发病率辖区每周报告的猴痘病例及预测。报告病例（黑色）以及使用VAR-Lasso（蓝色）、AR-Lasso（绿色）和朴素估计量（品红色）生成的预测。请参见补充章节中带预测区间的预测。

2.5 敏感性分析

每周猴痘病例数表现出随机噪声（图1）。平滑处理可降低这些高频波动的影响，并提高分析所需的信噪比。我们进行敏感性分析，以评估平滑窗口长度对每周猴痘数据及后续预测性能的影响。为确定最优窗口长度，使用不同长度的居中移动平均对输入时间序列进行平滑处理（不平滑，以及2周、3周、4周和5周窗口）。针对每个平滑窗口，为VAR-Lasso、AR-Lasso和朴¨素移动平均模型生成预测，并使用正斜率加权误差指标比较其性能。使用不同平滑长度生成的全部八个辖区42个预测周的斜率加权均方根误差和平均绝对误差见图7。

结果表明，平滑窗口的选择会显著影响预测准确性。基于原始数据或2周窗口的预测产生的误差最大，这可能反映了病例数中的噪声对可预测性的强烈影响。使用3周窗口时误差减小，并在使用4周移动平均时达到最小，此时均方根误差和平均绝对误差最低。然而，进一步平滑（5周窗口）后性能再次下降，表明过度平均会掩盖短期流行病动态。总体而言，采用4周平滑窗口的VAR-Lasso模型成为最优预处理选择，产生的误差最低。该平滑配置有效平衡了降噪与流行病信号的保留，因此被用于最终预测模型。

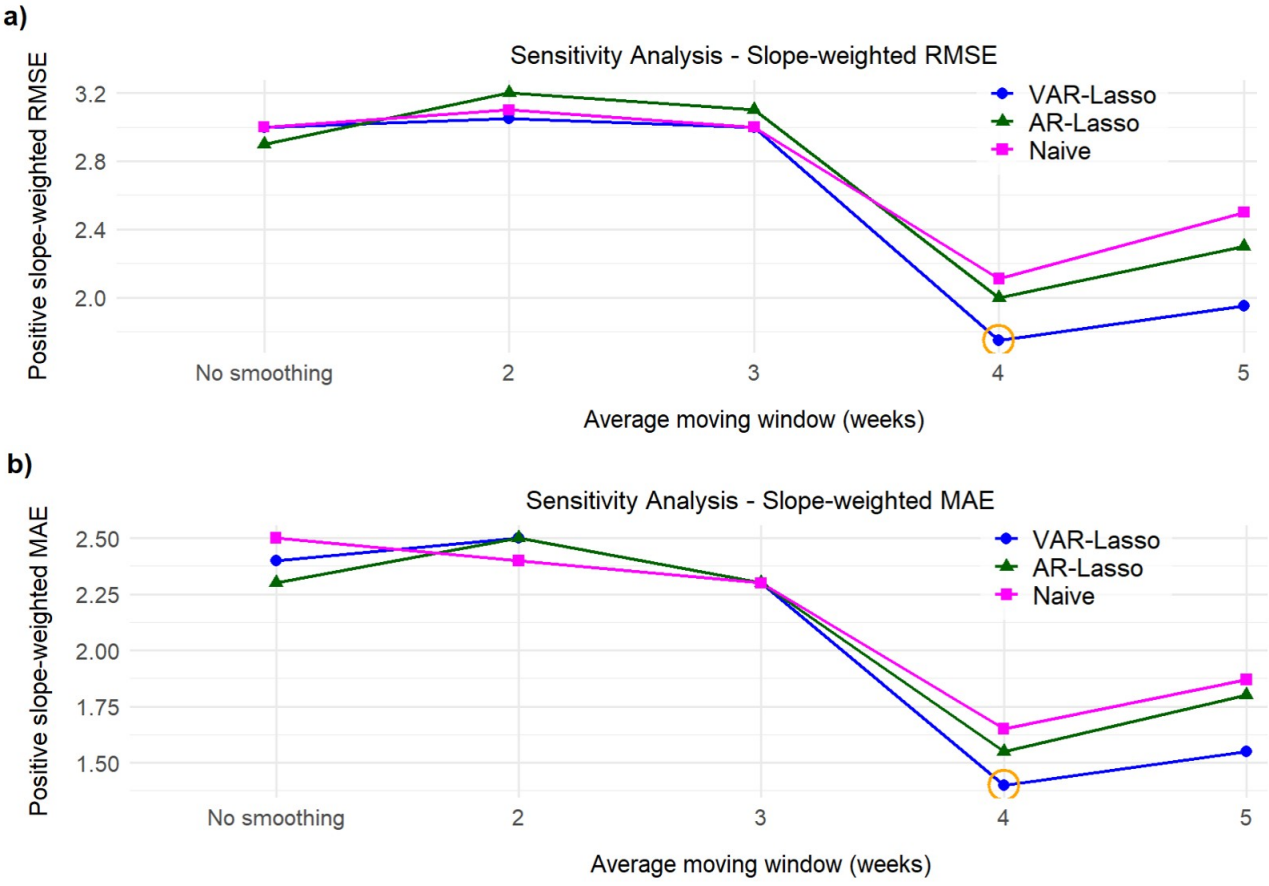

图 7：通过正斜率加权均方根误差 (a) 和正斜率加权平均绝对误差 (b) 衡量的平滑窗口时长对预测性能的敏感性分析。我们开展了敏感性分析，以评估不同平滑窗口（不进行平滑处理，以及 2、3、4 和 5 周窗口）对所比较的三种模型的预测性能的影响。误差最低的预测模型——采用 4 周移动平均窗口的向量自回归-Lasso——以橙色圆圈标出。

3 讨论

我们开发了一种新型 VAR-Lasso 猴痘预测模型，用于对美国八个高发病率辖区的每周猴痘病例数进行提前两周预测。与广泛使用的 AR 方法和朴素移动平均基线相比，VAR-Lasso模型的预测性能更优。

本研究展示了正则化多元时间序列方法在传染病预测中的一种新应用。向量自回归模型为捕捉辖区之间的动态相互依赖关系提供了自然的框架，但随着地点和滞后项数量增加，这类模型容易出现参数过多的问题，在时间跨度相对较短的疫情时间序列中尤其如此。采用 Lasso正则化可通过将许多系数压缩为零来应对这一挑战，从而减少过拟合，同时自动筛选出有信息量的滞后预测变量。由此得到的模型结构稀疏且易于解释，不仅提高了预测性能，还揭示了关键的跨辖区传播路径。

VAR-Lasso模型的表现优于更常见的 AR-Lasso 和朴素方法，这体现了纳入跨辖区依赖关系的价值。许多既往猴痘预测研究仅采用单变量时间序列方法，例如自回归（AR）和 ARIMA模型。这些方法能够捕捉单个辖区内的时间自相关，却无法纳入跨辖区相互作用。例如，Munir等人（2024）使用 ARIMA模型预测了2022年全球疫情期间受影响严重的10个国家的猴痘病例；Bakare等人（2025）则采用类似框架，对尼日利亚的猴痘病例和死亡率进行长期预测。虽然这些方法能够对单个地点作出合理预测，但在识别或预判由输入性传播或地理耦合动态驱动的传播方面存在固有局限。相比之下，本研究采用的VAR-Lasso框架利用辖区之间的多元依赖关系，使其能够纳入上游病例趋势，并识别有信息量的交叉滞后预测变量。这是对现有猴痘预测研究在方法上的拓展，也凸显了多元正则化模型在区域连通性对传播模式具有实质影响的情境中的附加价值。

VAR-Lasso模型识别出的滞后预测变量与独立系统发育证据相吻合，这为我们的建模方法提供了关键的外部验证。具体而言，我们将VAR-Lasso模型统计推断出的跨辖区效应与猴痘病毒实际传入SDC的基因输入事件联系起来；其中，从伊利诺伊州传入的23–24周长滞后期，与SDC优势簇的系统发育重建结果相呼应。这表明，稀疏多变量模型能够在本地疫情暴发达到高峰之前识别感染的上游来源。这些模型能够捕捉到经基因组流行病学独立支持的跨辖区依赖，凸显了它们在提供猴痘辖区间传播动态早期统计信号方面的潜力。

这项研究的另一项贡献是，VAR-Lasso模型能够提前两周针对病例上升发出早期预警信号。从公共卫生角度来看，提前预测病例增加的能力十分宝贵。CSDHHSA的卫生官员指出，此类早期预警预测有助于确定针对性宣传、外展工作和

疫苗接种宣传活动的开展时机，从而更有效地分配有限资源。

预测评估指标通常对所有时间点一视同仁，而不考虑底层过程是稳定还是快速变化。然而，某些流行病学特征比其他特征更具影响力，因此在评估预测性能时，可以相应地赋予不同权重（Tabataba 等人，2017；Armstrong，2001）。在发病率上升期间，公共卫生行动往往最为迫切，此时疫苗接种和健康教育宣传活动尤为关键。使用斜率加权误差指标可以通过优先评估传播加剧期间的预测性能来应对这一问题。

本研究存在若干重要局限。首先，鉴于猴痘近期才在美国出现，历史数据有限，限制了模型学习长期季节性或行为模式的能力；未来研究可从增加监测数据年份中获益。其次，猴痘仍存在漏诊和漏报问题，且这些问题在不同辖区之间可能呈现系统性差异，从而可能使病例数和推断出的跨辖区关系产生偏差。第三，模型仅依赖报告病例数据，未纳入其他可能有用的信号，例如流动模式、性网络数据、疫苗接种覆盖率或在线数据；这些信息可能提升预测性能，并有助于解释跨辖区依赖。第四，尽管系统发育数据可提供有价值的外部验证，但患者层面的基因组测序数据仅限于圣迭戈县。最后，尽管VAR-Lasso模型在大多数观测值和辖区中的表现均优于两个基准模型，但其在佛罗里达州的表现较差，在那里，AR-Lasso和朴素模型的预测更准确。这种差异可能反映出某些辖区的发病率更多地受到本地传播的驱动。在这种情况下，仅依赖辖区自身历史数据的单变量模型或许就已足够。

尽管存在这些局限性，我们的研究结果表明，整合跨辖区病例信息的稀疏多变量时间序列模型，能够显著加强对美国猴痘活动的早期检测和预测。VAR-Lasso模型能够识别与系统发育证据相符的长滞后预测因子，这凸显了其预测由远距离地理引入所致疫情暴发的潜力。未来一个有前景的方向是整合新的数据源，以提高预测准确度，并进一步支持 Resilient Shield 和 CSDHHSA 等公共卫生合作伙伴关系在圣迭戈县的工作。通过为公共卫生干预措施提供可采取行动的提前量，这种方法展现了将多变量统计建模与流行病学和基因组数据相结合，从而改进有针对性的公共卫生行动的价值。

4 方法

4.1 数据

每周猴痘病例数数据来自美国疾病控制与预防中心的国家监测系统，统计时间为2023年1月1日至2024年11月24日（可获取的最新数据）。我们选取了2023年和2024年猴痘累计发病率最高的八个美国辖区，这些辖区报告的病例数合计约占这两年美国报告病例总数的55%。这

八个入选辖区按2023–2024年期间病例总数从高到低依次为：纽约市（市）、得克萨斯州（州）、洛杉矶县（县）、佛罗里达州（州）、伊利诺伊州（州）、佐治亚州（州）、圣迭戈县（县）和华盛顿州（州）。

4.2 描述性统计——数据

我们对病例总数和每周病例数进行汇总并可视化，以描述时间趋势以及各辖区对全国发病率的相对贡献。报告的汇总统计包括每年的病例总数，以及各辖区病例数占美国病例总数的比例。

4.3 时间序列预处理：平滑处理与差分

为减少高频波动并提高信噪比，使用居中的 4 周移动平均对每周病例数进行平滑处理。对于每个辖区，在模型拟合中以平滑后的序列替代原始序列。我们进行了敏感性分析，以评估不同平滑窗口（2 周、3 周、4 周和 5 周窗口）的影响。还使用增广迪基–富勒（ADF）检验（迪基和富勒，1979）测试了各辖区病例时间序列的平稳性，因为拟合自回归模型需要满足平稳性。为实现平稳性，对每个时间序列进行一次差分。

4.4 向量自回归（VAR）模型

为对每周猴痘病例数的时间动态和跨辖区依赖进行建模，我们按照上述方法对处理后的时间序列拟合 VAR 模型。令 K 表示所建模的时间序列数量；在本研究中，$K = 8$，对应于所研究的八个美国辖区。令 $t \in \{1, 2, \dots\}$ 表示时间序列中的周数。对于每周 t，我们定义一个 K 维向量 $\mathbf{y}_t$，表示观测到的每周病例数：

$$\mathbf{y}_t = (y_{1t}, y_{2t}, \dots, y_{Kt})^\top,$$

其中，y_{kt} 表示辖区 k 在第 t周的猴痘病例数，$\top$ 表示该向量的转置。完整的观测时间序列可表示为以下序列：

$$\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots\}.$$

令 $\widetilde{y_{kt}}$，$\widetilde{\mathbf{y}}_t$，且 $\mathbf{Y}$ 表示平滑后的数据。滞后阶数为 p 的向量自回归模型将每周平滑后的数据 $\widetilde{\mathbf{y}}_t$ 表示为最近的 p 个前期每周观测值、截距项及随机误差项的线性函数：

$$\widetilde{\mathbf{y}}_t = \mathbf{c} + \Phi_1 \widetilde{\mathbf{y}_{t-1}} + \Phi_2 \widetilde{\mathbf{y}_{t-2}} + \cdots + \Phi_p \widetilde{\mathbf{y}_{t-p}} + \boldsymbol{\epsilon}_t.$$

其中，$\mathbf{c}$是一个 K维截距项向量，$\Phi_1, \dots, \Phi_p$是 $K \times K$个自回归系数矩阵，用于捕捉过去几周的病例数对当前周的影响，而 $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \Sigma)$是一个随机误差项向量，其协方差矩阵为 Σ。在我们的分析中，我们将滞后阶数设为 $p = 25$周。这一选择依据初步探索

性相关性分析作出，分析结果表明，各辖区每周病例数之间存在显著的时间依赖关系，最长可达25周。通过纳入这些滞后项，向量自回归模型能够捕捉猴痘在各辖区间传播的时间依赖关系。

4.5 向量自回归模型的 Lasso 正则化

随着辖区数量（K）和滞后阶数（p）的增加，向量自回归模型中的参数数量会迅速增长，从而面临过拟合的风险。为解决这一问题，我们采用 Lasso，这是一种兼具变量收缩和变量选择功能的正则化方法。Tibshirani 于 1996年 最初将 Lasso 引入线性回归。Lasso 在模型系数上引入 L1 惩罚项，根据惩罚项 λ 的取值，将系数收缩至零附近，并使其中一些系数恰好变为零。这实际上在估计模型系数的同时完成了变量选择。在向量自回归模型中，这使模型能够识别各辖区过去哪些滞后项或前几周的病例数对预测当前病例数最为重要。遵循 Hsu 等人（2008年）的做法，我们将第 t 周的滞后协变量向量定义为： K 辖区

$$\mathbf{X}_t = (1, \widetilde{\mathbf{y}_{t-1}}^\top, \widetilde{\mathbf{y}_{t-2}}^\top, \ldots, \widetilde{\mathbf{y}_{t-p}}^\top)^\top.$$

设计矩阵由所有周对应的向量排列而成： :

$$\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top,$$

wh ˜ich，其中包含与 $\boldsymbol{Y}$ 中的响应变量相对应的所有滞后预测变量。L et

$$\boldsymbol{\beta} = \mathrm{vec}(\mathbf{c}, \Phi_1, \ldots, \Phi_p)$$

˜表示回归形式中的系数向量。Lasso 估计量 $\boldsymbol{\beta}$ 通过最小化惩罚目标函数得到

$$Q(\boldsymbol{\beta}) = (\widetilde{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta})^\top (\widetilde{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_j |\beta_j|,$$

其中， λ 是用于控制收缩程度的 L1 惩罚项。当 L1 惩罚项 λ 较小时，Lasso 估计值会退化为普通最小二乘法（OLS）估计值。当 λ取值较大时，系数会向零收缩，其中一些系数会被精确地设为零，从而得到稀疏模型。在我们的分析中，Lasso正则化使我们能够在控制过拟合的同时，识别出美国各辖区最相关的过往周数值。按照 Tibshirani（1996）的建议，我们使用交叉验证选择调优参数 λ。

4.6 模型拟合与滚动预测设计

我们针对八个选定辖区在2024年1月至11月期间的全部42个预测周，采用滚动预测设计。对于每个预测周，VAR-Lasso模型（滞后 $= 25$周）均使用截至该预测周前两周、从2023年1月开始的所有既往数据进行拟合。VAR-Lasso模型通过R中的BigVAR软件包实现（Nicholson等，2025）。建模流程如下：

1. 采用4周居中移动平均对时间序列进行平滑处理并减少噪声 2. 使用增广迪基–富勒检验检验每个辖区序列的平稳性，并对序列进行一次差分 3. 使用截至预测周前两周的所有既往数据拟合VAR-Lasso模型（滞后 $= 25$周） 4. 生成提前两周的预测 5. 通过逆向差分将预测结果转换回原始尺度。

4.7 基准模型：AR-Lasso 和朴素估计器

为评估向量自回归–Lasso模型的预测性能，我们将其预测结果与两种基准方法进行比较：单变量自回归Lasso（AR–Lasso）模型和朴素移动平均估计器。AR–Lasso模型是一种采用Lasso正则化的单变量自回归模型，通过交叉验证选择有信息量的滞后项。该AR-Lasso模型使用R软件包sparseVAR实现（Zeng，2023年）。朴素基准方法通过将前四周观察到的平均病例数作为预测值，预测未来两周；在流感样疾病预测中也使用了类似的比较方法（Yang 等人，2017年）。

4.8 斜率加权性能指标

标准预测指标在评估预测准确性时，将所有时间点一视同仁，而不会优先考虑流行病学相关特征（Tabataba 等人，2017年；Armstrong，2001年）。在公共卫生领域，在预测评估中可能会对被认为影响更大的特征赋予更高权重（Tabataba 等人，2017年）。发病率上升的时期是疫情时间序列中一种流行病学相关时期，而在这些时期，采取公共卫生行动往往尤为迫切。为体现这一优先级，我们引入斜率加权误差指标，根据逐周正增长的幅度（即观察到的发病率正斜率的幅度）对预测误差加权。这样一来，传播增加的时期在评估中会产生更大的影响，从而优先考量模型在病例增长时间窗口内的表现。下文将正式定义用于评估预测准确性的斜率加权指标。令$\hat{}y_t$ 表示时间 t的预测值。观察到的发病率斜率定义为

$$\Delta y_t^{\text{pos}} = \max(\widetilde{y_t} - \widetilde{y_{t-1}}, 0), \tag{1}$$

其仅捕捉正向变化。斜率加权误差指标如下：

- 斜率加权均方根误差（RMSE）为

$$\text{Slope-weighted RMSE} = \sqrt{\frac{\sum_t (\hat{y}_t - y_t)^2 \cdot \Delta y_t^{\text{pos}}}{\sum_t \Delta y_t^{\text{pos}}}} \tag{2}$$

- 斜率加权平均绝对误差（MAE）为

$$\text{Slope-weighted MAE} = \frac{\sum_t |\hat{y}_t - y_t| \cdot \Delta y_t^{\text{pos}}}{\sum_t \Delta y_t^{\text{pos}}} \tag{3}$$

- 斜率加权偏差为

$$\text{Slope-weighted Bias} = \frac{\sum_t (\hat{y}_t - y_t) \cdot \Delta y_t^{\text{pos}}}{\sum_t \Delta y_t^{\text{pos}}} \tag{4}$$

4.9 系统发育分析

为刻画传入 SDC 的猴痘病毒谱系的进化历史，我们汇编并去除了 2023年和 2024年采样、并提交至 GenBank 和/或 GISAID 的猴痘病毒基因组序列比对中的重复项（n=2571 个基因组）。我们使用 SQUIRREL 对基因组进行序列比对，并采用 GTR+CAT20 替换模型，通过 FASTTREE2 推断最大似然树（Price 等，2010）。我们使用 TreeTime v0.11.4 中的迁移模型，识别出传入 SDC 的不同事件，这些事件由 29 个猴痘病毒基因组代表（Sagulenko 等，2018）。我们使用 BEASTX v10.5.0（Baele 等，2025），对五个猴痘病毒簇进行了贝叶斯系统地理学推断，其中每个簇至少包含一个在 SDC 采样的猴痘病毒基因组。对于每个簇，我们在 GTR+F 替换模型、严格分子钟（先验：中位数 $= 6 \times 10^{-5}$次替换/位点/年；标准差 $= 2 \times 10^{-5}$）（Pekar 等，2025）、SkyGrid 合并先验（Baele 等，2020）以及非对称离散性状模型的设定下，运行了两条独立链，每条链运行 1000 万至 4000 万代，以推断系统发育树中祖先的地理位置。我们使用 TreeSwift（Moshiri，2020），根据后验概率树分布汇总到达 SDC 之前的地理位置。

4.10 代码可用性

用于运行模型、生成预测、评估性能和创建可视化的代码，可在公共 GitHub 仓库中获取：`https://github.com/hanmacrad2/Mpox_VAR_Lasso_Automate`。

参考文献

Armstrong, J. S.（2001）。评估预测方法。在预测原理：研究人员和从业者手册中，第443–472页。Springer。

Baele, G.、Gill, M. S.、Lemey, P.和Suchard, M. A.（2020年）。在贝叶斯系统发育学框架中，采用哈密顿蒙特卡洛抽样与Skygrid溯祖模型估计历史种群动态。*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.等（2025年）。用于贝叶斯系统发育、系统地理和系统动力学推断的Beast X。*Nature Methods*，第1–4页。

Bakare, E. A.、Mogbojuri, O. A.、Oniyelu, D. O.、Abidemi, A.、Daniel, D. O.、Olasupo, I. I.、Osikoya, S. A.、Nwana, A. O.、Olorunfemi, R. D.和Olagbami, S. O.（2025年）。尼日利亚猴痘发病率与死亡率的时间序列建模与预测。*BMCInfectious Diseases*，25(1):794。Brand, S. P.、Cavallaro, M.、Cumming, F.、Turner, C.、Florence, I.、Blomquist, P.、Hilton, J.、Guzman-Rincon, L. M.、House, T.、Nokes, D. J.等（2023年）。疫苗接种与公众意识在英国猴痘发病率预测中的作用。*NatureCommunications*，14(1):4100。加利福尼亚州公共卫生部（2025年）。加利福尼亚州内I型分支猴痘的社区传播。
`https://www.cdph.ca.gov/Programs/OPA/Pages/CAHAN/Community-Spread-of-Clade-I-Mpox-Within-California.aspx`。访问日期：2025-10-17。美国疾病控制与预防中心（2025年）。美国病例趋势：II支猴痘。
`https://www.cdc.gov/monkeypox/data-research/cases/index.html`。访问日期：2025-11-17。Delaney, K. P.（2022年）。男同性恋者、双性恋者及其他男男性行为者为预防猴痘病毒传播而采取的策略——美国，2022年8月。*MMWR.*《发病率和死亡率周报》，71。Dickey, D. A.和Fuller, W. A.（1979年）。含单位根的自回归时间序列估计量的分布。《美国统计学会会刊》，74(366a):427–431。Hsu, N.-J.、Hung, H.-L.和Chang, Y.-M.（2008年）。使用Lasso进行向量自回归过程的子集选择。*Computational Statistics & Data Analysis*，52(7):3645–3657。Laurenson-Schafer, H.、Sklenovsk´a, N.、Hoxha, A.、Kerr, S. M.、Ndumbi, P.、Fitzner, J.、Almiron, M.、de Sousa, L. A.、Briand, S.、Cenciarelli, O.等（2023年）。首次猴痘全球暴发的描述：全球监测数据分析。*The Lancet GlobalHealth*，11(7):e1012–e1023。Mitj`a, O.、Ogoina, D.、Titanji, B. K.、Galvan, C.、Muyembe, J.-J.、Marks, M.和Orkin, C. M.（2023年）。猴痘。*The Lancet*，401(10370):60–74。

Moshiri, N.（2020年）。TreeSwift：大规模可扩展的Python树软件包。*SoftwareX*, 11:100436。Munir, T.、Khan, M.、Cheema, S. A.、Khan, F.、Usmani, A.和Nazir, M.（2024年）。10个主要受影响国家猴痘疫情趋势的时间序列分析与短期预测。*BMC Infectious Diseases*, 24(1):16。Nicholson, W.、Bien, J.、Matteson, D.和Wilms, I.（2025年）。*BigVAR*：含外生变量的稀疏向量自回归建模工具。R软件包版本1.0.0, https://cran.r-project.org/web/packages/BigVAR/vignettes/BigVAR.html。Overton, C. E.、Abbott, S.、Christie, R.、Cumming, F.、Day, J.、Jones, O.、Paton, R.、Turner, C.和Ward, T.（2023年）。英格兰2022年猴痘疫情的临近预测。*PLoS*计算生物学，19(9):e1011463。Owens, L. E.（2023年）。美国猴痘高风险人群的Jynneos疫苗接种覆盖率，2022年5月22日至2023年1月31日。*MMWR.*《发病率和死亡率周报》，72。Pekar, J. E.、Wang, Y.、Wang, J. C.、Shao, Y.、Taki, F.、Forgione, L. A.、Amin, H.、Clabby, T.、Johnson, K.、Torian, L. V.等（2025年）。纽约市2022年猴痘流行的传播动力学。《自然医学》，第1–10页。Pischel, L.、Martini, B. A.、Yu, N.、Cacesse, D.、Tracy, M.、Kharbanda, K.、Ahmed, N.、Patel, K. M.、Grimshaw, A. A.、Malik, A. A.等（2024年）。第三代猴痘疫苗针对猴痘及疾病严重程度的疫苗有效性：系统综述与荟萃分析。《疫苗》，42(25):126053。Price, M. N.、Dehal, P. S.和Arkin, A. P.（2010年）。FastTree 2——大型序列比对的近似最大似然树。*PLOS ONE*, 5(3):e9490。Sagulenko, P.、Puller, V.和Neher, R. A.（2018年）。TreeTime：最大似然系统发育动力学分析。《病毒进化》，4(1):vex042。Tabataba, F. S.、Chakraborty, P.、Ramakrishnan, N.、Venkatramanan, S.、Chen, J.、Lewis, B.和Marathe, M.（2017年）。疫情预测评估框架。*BMC Infectious Diseases*，17(1):345。Tibshirani, R.（1996年）。经由Lasso的回归收缩与选择。《英国皇家统计学会会刊*B*辑：统计方法学》，58(1):267–288。Yang, S.、Santillana, M.、Brownstein, J. S.、Gray, J.、Richardson, S.和Kou, S.（2017年）。利用电子健康记录与互联网搜索信息进行准确的流感预测。*BMC Infectious Diseases*，17(1):332。Zeng, D.（2023年）。*sparseVAR*：稀疏向量自回归建模。R文档。R软件包版本0.2.3。

作者贡献

HC、JOW、MB、EAS、SS、RG 和 NKM 构思了本研究。HC、JOW、RG 和 NKM 开发了统计模型。DV 和 RG 协助处理数据。HC 和 JW 开展了正式分析。HC、JOW、LR、RG 和 NKM 撰写了论文初稿。HC、JOW、EAS、MB、DV、RG、JW、LR、SS、RG 和 NKM 审阅了稿件，并参与了最终版本的完善。

致谢

作者衷心感谢来自病毒样本来源实验室和提交实验室的作者，他们通过 GISAID 生成并共享了本研究所依据的病毒基因组序列和元数据（补充表）。

资助

本项目得以开展，得益于美国疾病控制与预防中心疫情预测与分析中心提供的合作协议 CDC-RFA-FT-23-0069（拨款编号：1 NU38FT000006-01-00）。项目内容完全由作者负责，并不一定代表美国疾病控制与预防中心的官方观点。

补充材料

预测区间构造

两步预测区间采用应用于一阶差分时间序列的递归预测框架构建。对于 VAR-Lasso模型，我们使用差分训练数据拟合模型，并使用 BigVAR R软件包的 `predict`函数中的 `confint` 选项，获得 $t+1$ 的一步预测及预测区间。随后，我们将 $t+1$ 点预测值扩展至训练序列中，重新拟合模型，并生成 $t+2$ 的一步预测及预测区间。累计两步变化的方差计算如下：

$$\mathrm{Var}(\Delta y_{t+1} + \Delta y_{t+2}) = \mathrm{Var}(\Delta y_{t+1}) + \mathrm{Var}(\Delta y_{t+2} \mid \widehat{y_{t+1}}).$$

方差计算隐含地假设，在给定 $\widehat{y_{t+1}}$的条件下，时间步$t+1$和 $t+2$的预测误差相互独立。这种方法会传播不确定性，但可能产生反保守的区间，因为它既未考虑不同时间步误差之间的协方差，也未考虑模型选择带来的不确定性。Lasso等惩罚回归方法得出的预测区间，无法体现变量选择带来的不确定性、正则化造成的收缩偏差，也未纳入完整的参数估计不确定性（Nicholson等，2025）。

对于AR-Lasso模型，我们对每个辖区分别采用类似的两步程序。在每次预测时，我们使用R的 `arima` 函数，通过最大似然估计对差分序列拟合AR模型，然后使用模型内置的 `predict` 函数生成直接的两步超前预测（`n.ahead = 2`）。所得标准误通过自回归系数自然地纳入了一步和两步预测误差之间的相关性。我们使用这些标准误构造95%预测区间，再将其从差分尺度转换回水平值，即把预测变化量累加到最近一次观测值上。

未来的工作可以探索更为严谨的不确定性量化方法，包括自助重抽样、去偏Lasso或贝叶斯方法，以获得校准效果更佳的预测区间，并充分考虑惩罚估计程序中的全部不确定性。

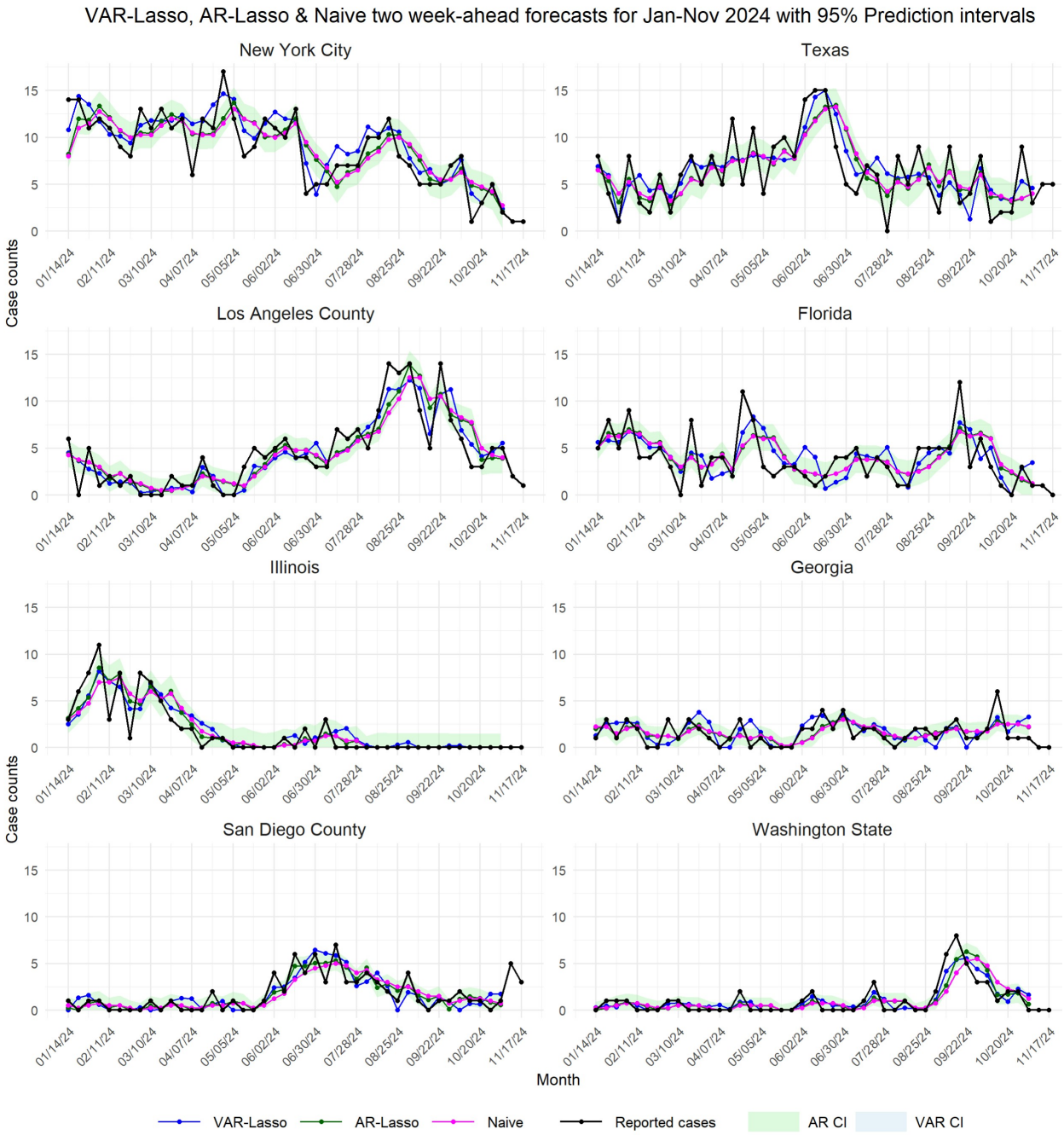

图 S1：2024年1月至11月美国8个高发病率辖区的预测值、95%预测区间及每周报告的猴痘病例。报告病例（黑色）、VAR-Lasso（蓝色）、VAR-Lasso预测区间（浅蓝色）、AR- Lasso（绿色）、AR-Lasso预测区间（浅绿色）和朴素估计量（洋红色）。

簇	Size（基因组）	圣迭戈病例（数量）	圣迭戈最近共同祖先时间：中位数（95% 最高后验密度区间）	起源地（后验概率）
A	433	4	2023.67 (2023.57, 2023.69)	加利福尼亚州 (0.962) 加利福尼亚州洛杉矶县 (0.032)
B	142	19	2023.69 (2023.57, 2023.75)	加利福尼亚州 (0.266) 加利福尼亚州洛杉矶县 (0.016) 伊利诺伊州 (0.717)
C	244	4	2024.43 (2024年.38, 2024年.47)	加利福尼亚州 (0.271) 加利福尼亚州洛杉矶县 (0.693) 纽约州-纽约市 (0.030)
D	102	1	N/A	加利福尼亚州 (1.000)
E	185	1	N/A	巴西 (0.030) 加拿大 (0.721) 伊利诺伊州 (0.022) 纽约州-纽约市 (0.175) 荷兰 (0.024) 葡萄牙 (0.020)

表 S1：圣迭戈县猴痘病例的系统发育簇特征。tMRCA，最近共同祖先时间；HPD，最高后验密度；CA，加利福尼亚州；IL，伊利诺伊州；CA-LAC，洛杉矶县；NY-NYC，纽约市。