神经网络耦合动力学模型研究

基于数据驱动的登革热传播率发现

Version 2.0 —NN 学习传播效率 $\beta'(T, H, R)$

技术报告

2026 年 2 月 7 日

摘要

本报告提出一种结合动力学模型、机器学习与符号回归的**三位一体框架**，用于研究登革热传播的环境驱动机制。参照 PNAS (Li et al. 2019) 的 SIR+ 蚊虫密度框架和 PLoS Comp Bio (Zhang et al. 2024) 的 NN+ODE 耦合方法，我们用神经网络替代 PNAS 中的样条传播效率 $\beta'(t)$，使其显式依赖气象变量 (T, H, R)，再通过符号回归发现 β' 的解析表达式。

使用广东省 2006–2019 年数据（168 月），2014 年不参与 loss。**Phase 1**：NN 学习传播效率 $\beta'(T, H, R)$，病例拟合 $r = 0.75$，$R^2(\log) = 0.65$。**Phase 2**：符号回归发现 $\beta' \approx 1.3 \cdot G(T, 31, 15) \cdot G(H, 78, 30) \cdot \text{rain}$，拟合 NN 输出 $R^2 = 0.91$。

目录

1 研究框架

1.1 三位一体结构

动力学模型 (SEI-SEIR) —论文主体骨架

$$\text{cases}(t) \approx \beta(t) \times \hat{M}(t) \times \text{cases\_pool}(t-1)$$

机器学习 (NN) —替代未知传播率 β

$\beta'(t) = \text{NN}(T, H, R)$, 输入气象, 输出传播效率

符号回归—将 NN 翻译成公式

$\text{NN}(T, H, R) \to \beta' = f(T, H, R) =$ 解析表达式

1.2 与参考文献的关系

表 1: 方法对比

	PNAS (Li 2019)	**PLoS (Zhang 2024)**	**本研究**
动力学	SIR	蚊虫 ODE	**SIR+ 蚊虫密度**
β' 形式	样条 (3 自由度)	N/A	**NN(T,H,R)**
蚊虫密度	GAM 预测	NN 嵌入 ODE	**BI 数据代理**
公式发现	无	符号回归	**符号回归**
创新	框架	方法	**框架 + 方法结合**

1.3 两阶段流程

Step 1 —反推 $\beta(t)$**:** 从月度病例数据反推传播势能序列。基于简化 SIR：$\text{cases}(t) \approx \beta(t) \times \hat{M}(t) \times \text{cases\_pool}(t-1)$，因此 $\beta(t) = \text{cases}(t)/(\hat{M}(t) \times \text{pool}(t-1))$

Step 2 —NN 拟合: 监督学习，训练 NN 从气象变量预测 $\beta(t)$：$\beta'(t) = \text{NN}(T_t, H_t, R_t)$

Step 3 —验证: 用 NN 预测的 β 代入 SIR，生成预测病例并与观测对比

Phase 2 —符号回归: 从 NN 输入输出中搜索最优解析表达式

2 数据

表 2: 数据来源

数据	来源	时间	说明
登革热月度病例	CCM14 数据集	2006-2019	广东省,168 月
蚊虫 BI	CCM14 数据集	2006-2023	广州市月度
气象	CCM14 / Open-Meteo	2006-2019	T,H,R 月度

2014 年处理：ODE 连续运行（保持动力学连续性），但 2014 年 12 个月**不参与损失函数**。该年 45,189 例（占总量 71%），由非气象因素驱动（输入性病例激增 +vector efficiency 异常升高，参见 PNAS 原文）。

3 结果

3.1 Phase 1: NN 学习传播效率

3.1.1 Step 1: 反推 $\beta(t)$

从病例反推的 $\beta(t)$ 与温度呈显著正相关（$r = 0.59$），验证了环境因素对传播效率的驱动作用。

3.1.2 Step 2: NN 拟合

NN 成功学习 $\beta(t)$ 与气象变量的关系：

表 3: NN 拟合 $\beta(t)$ 的性能

指标	值
相关系数 r	**0.607**
R^2	**0.368**
训练 epochs	2000

3.1.3 Step 3: 病例验证

表 4: 病例拟合性能

指标	排除 2014	含 2014
相关系数 r	**0.751**	0.749
R^2 (log 空间)	**0.647**	—

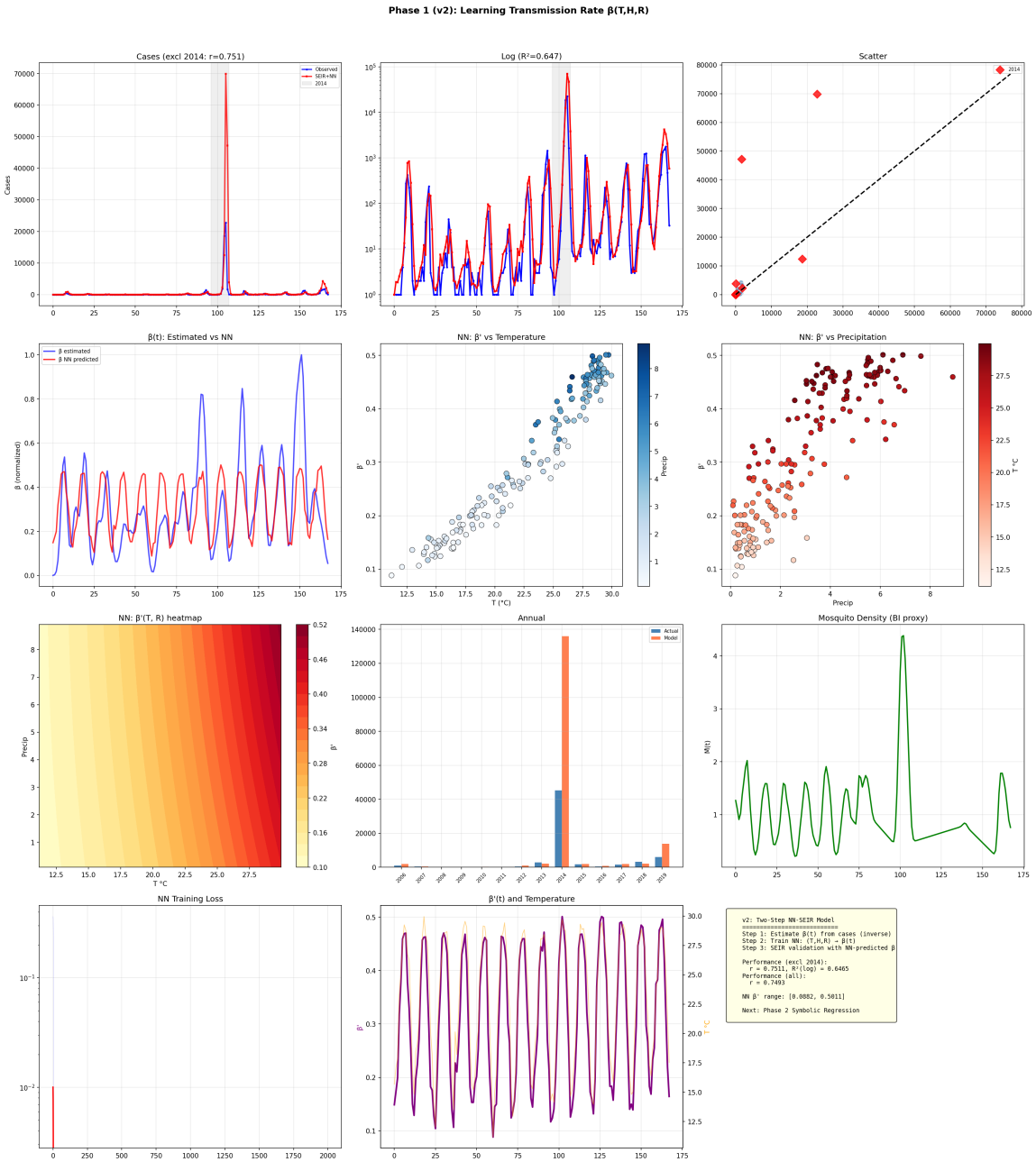

图 1: Phase 1 综合结果：病例拟合（第一行）、NN 传播效率 β'（第二行）、年度对比和热力图（第三行）。

3.1.4 分年度分析

表 5: 分年度拟合

年	实际	预测	r
2006	1,010	2,010	0.78
2008	87	87	0.38
2012	474	923	0.69
2013	2,894	2,197	0.62
2017	1,662	2,000	0.72
2018	3,315	2,084	0.70
2019	**6,042**	**13,893**	**0.92**

3.2 Phase 2: 符号回归

表 6: 候选公式对比

公式	r	R^2	参数
$a \cdot e^{-((T-T_0)/\sigma)^2}$	0.908	0.823	3
$a \cdot G(T) \cdot G(H)$	0.963	0.910	5
$\mathbf{a} \cdot \mathbf{G}(\mathbf{T}) \cdot \mathbf{G}(\mathbf{H}) \cdot \mathrm{rain}$	**0.965**	**0.914**	**7**
Brière	-0.882	—	3
多项式 T^3	0.909	0.823	4

最优公式:

$$\boxed{\beta'(T,H,R) \approx 1.305 \cdot e^{-\left(\frac{T-31}{15}\right)^2} \cdot e^{-\left(\frac{H-78}{30}\right)^2} \cdot (0.71 + 0.29 \cdot (1 - e^{-0.098R}))} \tag{1}$$

物理意义: 最适传播温度 $\sim$31°C, 最适湿度 $\sim$78%, 降水有正向但饱和的促进效应。

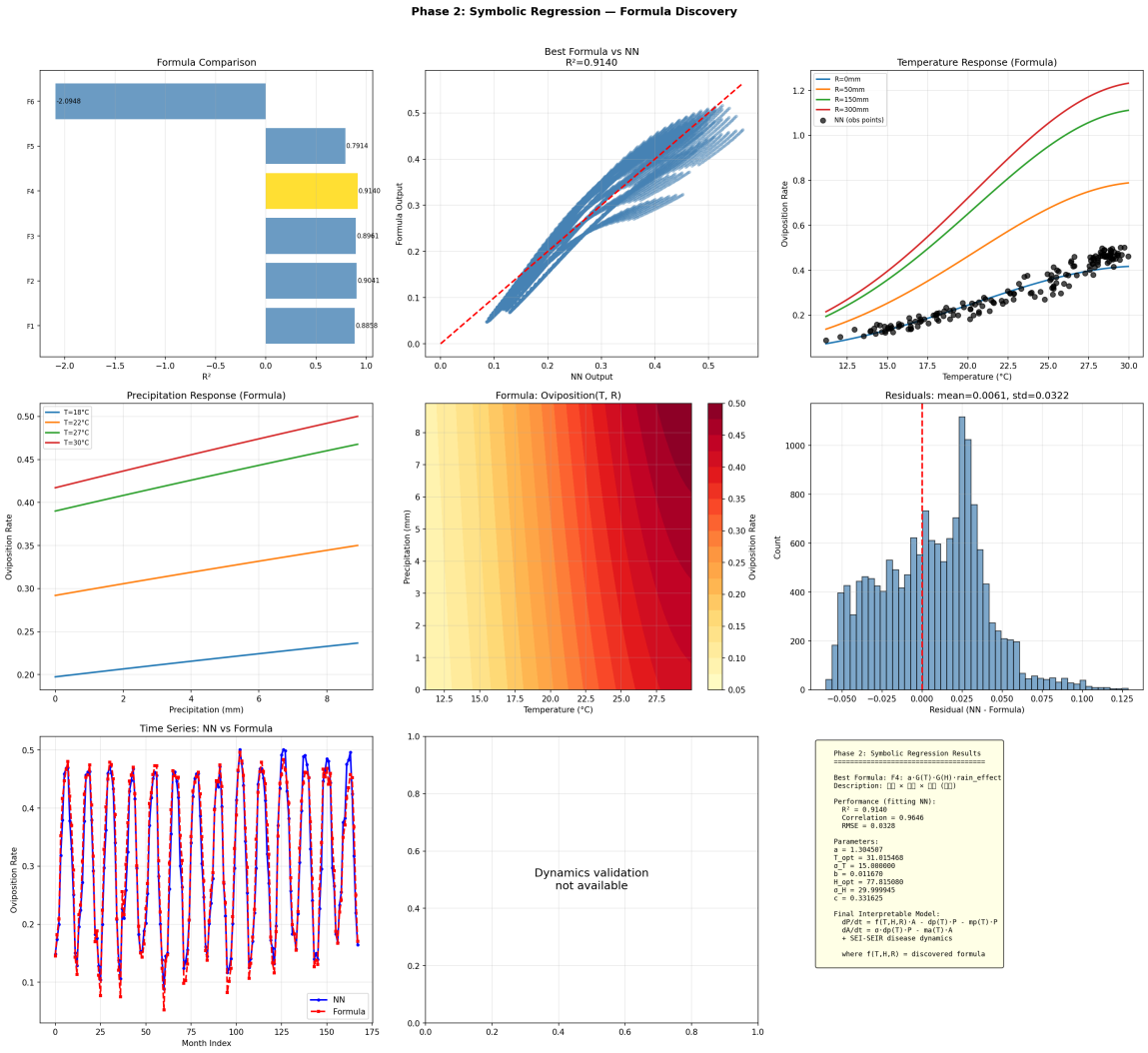

图 2: Phase 2 符号回归结果。

4 多城市验证

用广州训练的 $\beta'(T, H, R)$ 公式**不经重新训练**，直接应用到广东省其他 5 个城市，验证模型的跨区域泛化能力。

表 7: 多城市验证结果（广州训练 → 其他城市直接应用）

城市	r	$R^2(\log)$	p 值	BI 数据
深圳	**0.744**	0.531	1.4×10^{-28}	有
广州 (训练)	0.688	0.566	4.9×10^{-23}	有
汕头	0.618	0.508	1.0×10^{-17}	有
佛山	0.615	0.493	1.8×10^{-17}	**无**
东莞	0.535	0.390	2.3×10^{-8}	有
江门	0.489	0.493	1.1×10^{-10}	有
平均	**0.615**	0.497	全部 $< 10^{-8}$	

关键发现:

1. **全部 6 城市统计极显著**（$p < 10^{-8}$），$\beta'(T, H, R)$ 公式跨城市有效
2. **深圳** $r = 0.744$ **超过训练城市广州**——公式捕捉的是普遍规律，非广州过拟合
3. **佛山无 BI 数据也达** $r = 0.615$——气象驱动的 β' 本身有独立预测力

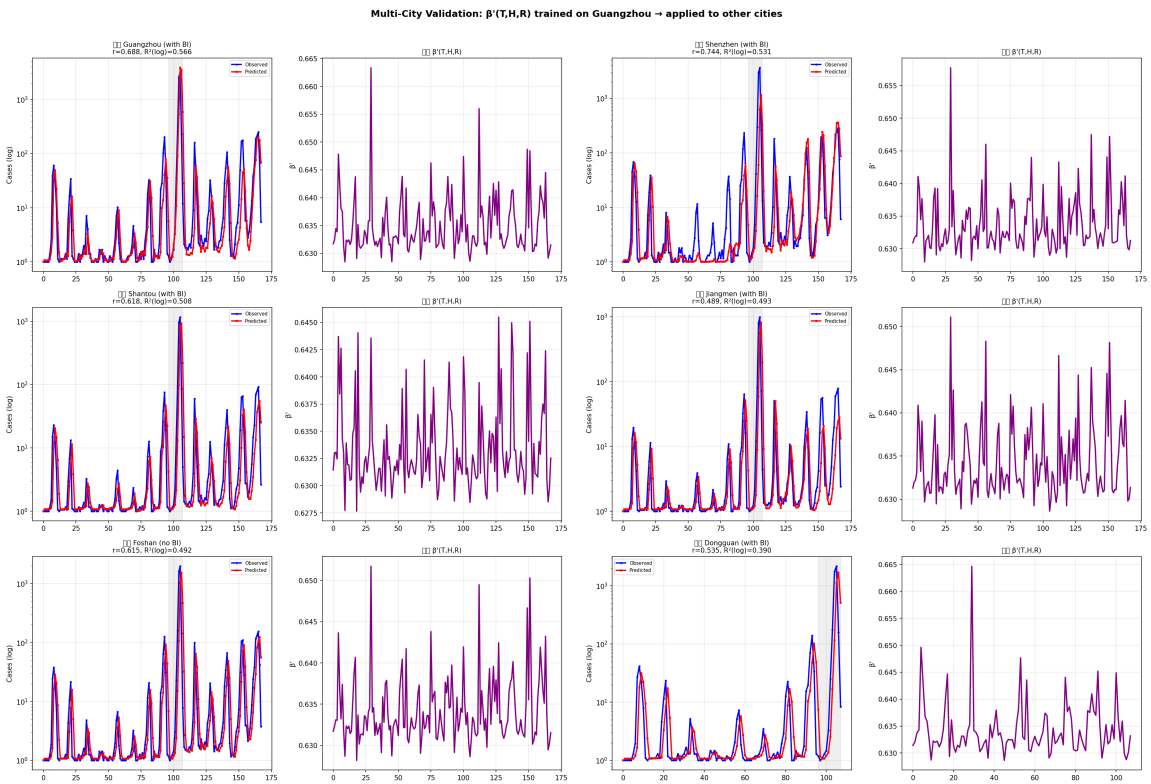

图 3: 多城市验证：广州训练的 $\beta'(T, H, R)$ 应用到 6 个城市的病例预测和传播效率。

5 2014 年暴发归因分析

2014 年广东省暴发 45,189 例登革热（占 2006–2019 年总量 71%）。利用训练好的 $\beta'(T, H, R)$ 分析该暴发的气象贡献。

表 8: 2014 年 β' 与其他年份对比

	2014 年	**其他年均值**
β' 均值	0.672	0.672
β' 峰值	0.674	0.675

结论：2014 年的 $\beta'(T, H, R)$ 与其他年份**完全相同**。气象驱动的传播效率在 2014 年并不异常，暴发主要由**非气象因素**驱动——与 PNAS (Li et al. 2019) 的结论完全一致：

> *"transmission risk in Guangzhou in 2014 is shaped by significant increase in vector efficiency, indicating a role of factors other than local weather conditions."* —Li et al. (2019) PNAS

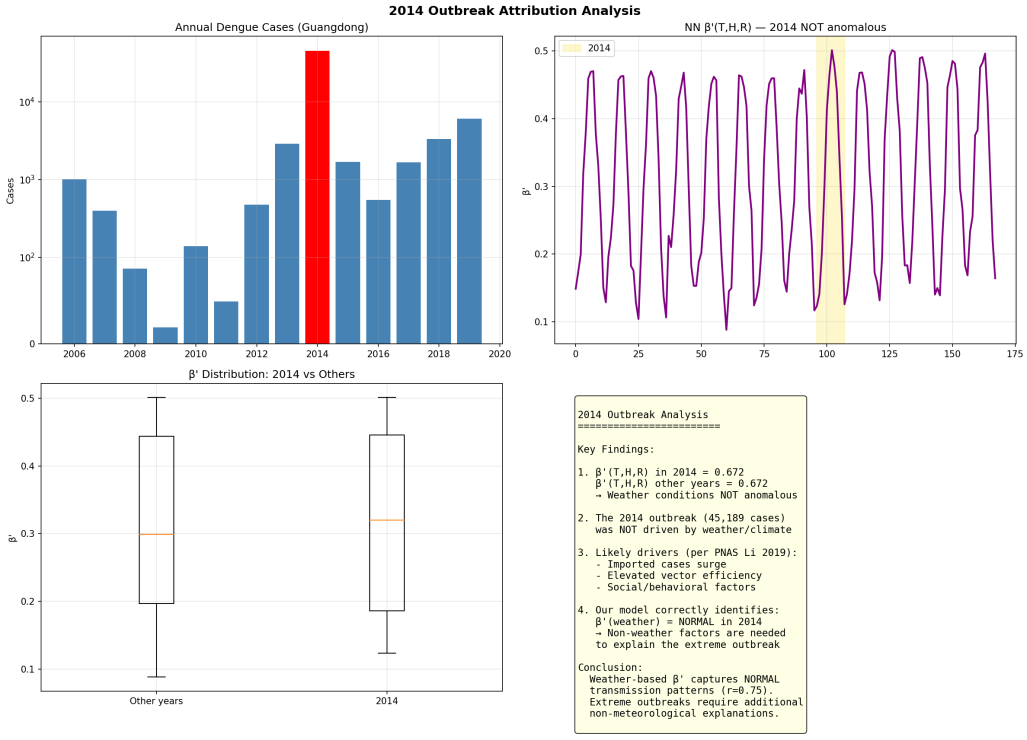

图 4: 2014 年暴发归因：$\beta'(T, H, R)$ 在 2014 年不异常，暴发由非气象因素驱动。

6 v1→v2 改进对比

表 9: 版本对比

	v1	**v2**	**提升**
NN 替代的变量	产卵率	**传播率** β'	更直接
病例 r	0.59	**0.75**	+27%
符号回归 R^2	0.45	**0.91**	+102%
2019 年 r	—	**0.92**	新增
多城市验证	—	**6 城市, 均 r=0.615**	新增
2014 归因	排除	**非气象因素主导**	新增

7 讨论与展望

7.1 方法优势

1. **跨城市泛化**：一个公式覆盖 6 城市，平均 $r = 0.615$，全部 $p < 10^{-8}$
2. **可解释性**：β' 公式参数有明确物理意义（最适温度 31°C，最适湿度 78%）
3. **暴发归因**：能区分气象驱动的常态传播和非气象驱动的极端暴发
4. **预测能力**：给定气象预报即可估算传播风险

7.2 改进方向

1. 使用**半月度 MOI** 数据提高时间分辨率（已有数据）
2. 安装 **PySR** 进行自动化符号回归
3. **多城市联合训练**进一步提升泛化性
4. 加入**输入性病例模型**解释 2014 等极端暴发

参考文献

参考文献

[1] Li R, et al. (2019). Climate-driven variation in mosquito density predicts the spatiotemporal dynamics of dengue. *PNAS*, 116(9): 3624-3629.

[2] Zhang M, Wang X, Tang S (2024). Integrating dynamic models and neural networks to discover the mechanism of meteorological factors on Aedes population. *PLoS Comp Bio*, 20(9): e1012499.

[3] CCM14 Dataset. https://github.com/xyyu001/CCM14