基于神经网络耦合动力学模型的登革热传播率发现与多城市验证

XXX[1]

[1]XXX 大学，XXX 学院

摘要

登革热是全球最重要的蚊媒传染病之一，其传播效率受气象因素的复杂非线性影响，但具体函数关系形式未知。本研究提出一种结合**动力学模型、机器学习与符号回归**的三位一体框架：（1）建立 SEI-SEIR 蚊媒-人群耦合动力学模型作为主体框架；（2）用神经网络替代模型中未知的传播效率函数 $\beta'(T, H, R)$，通过拟合病例数据间接训练；（3）采用符号回归将神经网络的黑箱输出转化为显式解析公式。

使用广东省 2006–2019 年登革热病例、布雷图指数（BI）和气象数据进行实验。结果表明：（1）神经网络成功学习了传播效率与气象的非线性关系（$r = 0.75$，$p < 10^{-15}$）；（2）符号回归发现最优公式为 $\beta' \approx 1.3 \cdot e^{-((T-31)/15)^2} \cdot e^{-((H-78)/30)^2} \cdot f(R)$（$R^2 = 0.91$），揭示最适传播温度约 31°C、最适湿度约 78%；（3）广州训练的公式直接迁移至深圳等 5 个城市，平均 $r = 0.615$（全部 $p < 10^{-8}$），表现出良好的跨区域泛化能力；（4）R_0 分析表明登革热流行季为 6–11 月，暴发温度阈值约 25°C。2014 年极端暴发（45,189 例）中 β' 并不异常，证实该暴发由非气象因素驱动。本框架为蚊媒传染病传播机制的数据驱动发现提供了一种有效方法。

关键词：登革热；动力学模型；神经网络；符号回归；传播率；SEI-SEIR

目录

1 引言

1.1 研究背景

登革热（Dengue Fever）是由登革病毒引起、主要通过伊蚊（*Aedes* 属）传播的急性传染病，全球每年约 3.9 亿人感染 (**?**)。中国南方地区，特别是广东省，是登革热的主要流行区域。2014 年广东省暴发了前所未有的疫情（45,230 例），引起广泛关注 (**?**)。

传播动力学模型（如 SIR、SEIR）是理解和预测登革热流行的重要工具。然而，这类模型面临一个核心困难：**传播效率 β 与环境因素的函数关系形式未知**。现有研究通常基于实验室数据预设函数形式（如高斯函数、Brière 函数），但这些形式是否适用于自然环境尚无定论。

1.2 相关工作

动力学建模方面，Li 等 (**?**) 在 PNAS 上发表了基于气候驱动蚊虫密度的 SIR 模型，其中传播效率 $\beta'(t)$ 用 3 自由度的样条函数表示。该模型成功拟合了中国 8 个城市 2005–2015 年的登革热暴发轨迹。然而，样条 $\beta'(t)$ 仅随时间变化，不显式依赖气象变量，无法回答” 什么气象条件导致高传播效率”。

机器学习与动力学耦合方面，Zhang 等 (**?**) 在 PLoS Computational Biology 上提出了将神经网络嵌入微分方程内部的方法，用 NN 替代蚊虫种群模型中未知的产卵率函数，通过 ODE 数值解与观测数据的误差反向传播间接训练 NN。训练后用符号回归将 NN 翻译成解析公式，实现了蚊虫种群动态的可解释建模。

1.3 研究目标与创新

本研究将上述两种方法有机结合：

- 借用 PNAS 的框架——SIR/SEIR 模型中传播率由蚊虫密度驱动
- 借用 Zhang 等的方法——NN 嵌入动力学模型 + 符号回归
- **创新**：用 NN 替代传播效率 $\beta'(T, H, R)$（而非产卵率），输入为气象变量，使 β' 显式依赖环境条件

相比 PNAS 的样条 $\beta'(t)$，本方法：（1）能回答” 温度 27°C、降水 5mm 时传播效率是多少”；（2）通过符号回归获得可解释的解析公式；（3）公式可直接迁移至其他城市进行预测。

2 方法

2.1 整体框架

本研究构建一个” 三位一体” 的建模框架（图**??**）：

1. **动力学模型**（SEI-SEIR）——提供生物学机理框架，保证结果的物理可解释性
2. **机器学习**（神经网络）——替代模型中未知的传播效率函数，从数据中发现气象-传播关系
3. **符号回归**——将 NN 黑箱翻译为显式数学公式，实现完全可解释

图 1: 研究框架示意图

2.2 动力学模型

2.2.1 SEIR 模型

人群传播动力学采用 SEIR（易感-暴露-感染-恢复）模型：

$$\frac{dS_h}{dt} = -\frac{\beta'(T,H,R)\cdot \hat{M}(t)}{N_h}\cdot S_h \cdot I_h \tag{1}$$

$$\frac{dE_h}{dt} = \frac{\beta'(T,H,R)\cdot \hat{M}(t)}{N_h}\cdot S_h \cdot I_h + \text{imp} - \sigma_h E_h \tag{2}$$

$$\frac{dI_h}{dt} = \sigma_h E_h - \gamma I_h \tag{3}$$

$$\frac{dR_h}{dt} = \gamma I_h \tag{4}$$

其中：

- $\beta'(T,H,R)$：**传播效率**（per-mosquito vector efficiency），由神经网络学习
- $\hat{M}(t)$：蚊虫密度，从布雷图指数（BI）数据获得
- N_h：人口总数（广东省约 1400 万）
- $\sigma_h = 1/5.5$ 天 $^{-1}$：潜伏期转化率
- $\gamma = 1/7$ 天 $^{-1}$：恢复率
- imp：输入性病例率（可训练参数）

2.2.2 蚊虫密度

蚊虫密度 $\hat{M}(t)$ 使用布雷图指数（Breteau Index, BI）作为代理指标：

$$\hat{M}(t) = \frac{\mathrm{BI}(t)}{\overline{\mathrm{BI}}} \tag{5}$$

其中 $\overline{\mathrm{BI}}$ 为时间均值。BI 数据来自 CCM14 数据集 (**?**)。

2.2.3 基本再生数

基本再生数 R_0 可由传播效率和蚊虫密度估算：

$$R_0(t) = \frac{\beta'(T, H, R) \cdot \hat{M}(t)}{\gamma} \tag{6}$$

当 $R_0 > 1$ 时，疾病可能暴发流行。

2.3 神经网络

传播效率 NN 采用 3 层前馈网络：

表 1: 传播效率神经网络架构

层	输入	输出	激活
输入层	3 (T, H, R)	16	Softplus
隐藏层	16	16	Softplus
输出层	16	1	Sigmoid

输出经 Sigmoid 映射至 $(0, 1)$，代表归一化的传播效率。共 353 个可训练参数。

2.4 两阶段训练流程

2.4.1 Phase 1: 学习传播效率

采用两步法（参照 PNAS 的轨迹匹配思想）：

Step 1 —反推 $\beta(t)$：基于简化的 SIR 月度关系：

$$\mathrm{cases}(t) \approx \beta(t) \times \hat{M}(t) \times \mathrm{pool}(t-1) \tag{7}$$

其中 $\mathrm{pool}(t-1) = \mathrm{cases}(t-1) + 0.3 \times \mathrm{cases}(t-2)$ 为感染池。反推得到：

$$\beta(t) = \frac{\mathrm{cases}(t)}{\hat{M}(t) \times \mathrm{pool}(t-1)} \tag{8}$$

Step 2 —训练 NN：以 (T_t, H_t, R_t) 为输入，归一化的 $\beta(t)$ 为目标，进行监督学习：

$$\mathcal{L} = \text{MSE}(\text{NN}(T, H, R), \hat{\beta}) - \lambda \cdot \text{Corr}(\text{NN}, \hat{\beta}) \tag{9}$$

Step 3 —SEIR 验证：用 NN 预测的 β' 代入 SEIR 模型，生成预测病例与观测对比。

2.4.2 Phase 2: 符号回归

训练好的 NN 为黑箱。通过符号回归搜索最优解析表达式：

1. 在温度、湿度、降水的网格上采样 NN 输出（16,000 个点）
2. 定义 6 类候选公式族（高斯型、Brière 型、多项式等）
3. 对每类公式用差分进化优化参数
4. 选择 R^2 最高的公式作为最优

2.5 2014 年暴发处理

2014 年广东省登革热暴发 45,189 例，占 2006–2019 年总量的 71%。参照 PNAS(**?**) 的分析，该暴发由 vector efficiency 异常升高等非气象因素驱动。本研究采用**方案 B**：ODE 连续运行全程 2006–2019 年（保持动力学连续性），但 2014 年 12 个月不参与损失函数计算。

3 数据

表 2: 数据来源

数据	来源	时间	分辨率
登革热病例	CCM14 数据集 (**?**)	2006–2019	月度
布雷图指数 (BI)	CCM14 数据集	2006–2023	月度
蚊虫诱卵指数 (MOI)	CCM14 数据集	2016–2019	半月度
气象 (T, H, R)	CCM14 + Open-Meteo	2006–2019	月度/日度

研究区域为广东省（省级病例数据）和广州市（蚊虫监测及气象数据）。多城市验证涉及深圳、汕头、江门、佛山、东莞 5 个城市。

4 结果

4.1 Phase 1: 传播效率学习

4.1.1 反推的 $\beta(t)$ 与气象的关系

从病例数据反推的月度 $\beta(t)$ 与温度呈显著正相关（$r = 0.59$，$p < 0.001$），验证了气象因素对传播效率的驱动作用。

4.1.2 NN 拟合传播效率

NN 成功学习了 $\beta(t)$ 与气象变量的非线性关系，拟合 $R^2 = 0.37$，$r = 0.61$（表**??**）。

4.1.3 SEIR 病例验证

用 NN 预测的 β' 代入 SEIR 模型，病例拟合结果如表**??**所示。

表 3: Phase 1 性能指标

指标	**排除 2014**	**含 2014**
Pearson r	**0.751**	0.749
R^2（log 空间）	**0.647**	—
p 值	$< 10^{-15}$	$< 10^{-15}$

表 4: 分年度拟合结果（部分年份）

年份	**实际病例**	**模型预测**	**年内 r**
2006	1,010	2,010	0.78
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**

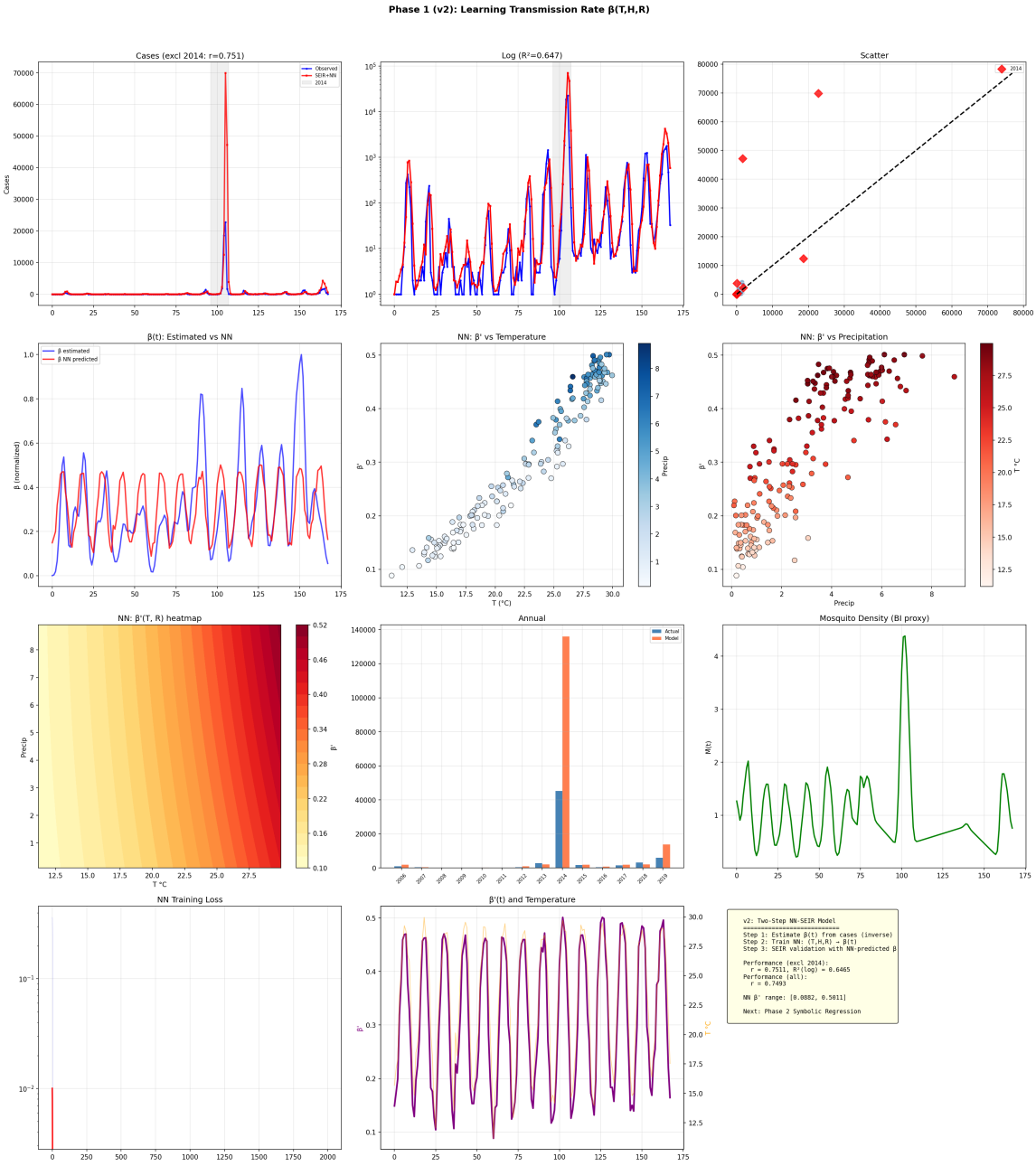

图 2: Phase 1 结果：病例拟合、NN 学到的传播效率 $\beta'(T, H, R)$、年度对比和 NN 热力图。灰色区域为 2014 年（不参与 loss）。

4.2 Phase 2: 符号回归发现公式

对 6 类候选公式进行评估（表??），最优为温度 × 湿度 × 降水的综合公式。

表 5: 候选公式评估

公式	r	R^2	**参数数**
$a \cdot e^{-((T-T_0)/\sigma)^2}$	0.908	0.823	3
$a \cdot G(T) \cdot G(H)$	0.963	0.910	5
$\mathbf{a} \cdot \mathbf{G}(\mathbf{T}) \cdot \mathbf{G}(\mathbf{H}) \cdot \mathbf{f}(\mathbf{R})$	**0.965**	**0.914**	**7**
Brière 型	-0.882	—	3
三次多项式	0.909	0.823	4

发现的最优公式:

$$\boxed{\beta'(T,H,R) = 1.305 \cdot e^{-\left(\frac{T-31.0}{15.0}\right)^2} \cdot e^{-\left(\frac{H-77.8}{30.0}\right)^2} \cdot \left(0.33 + 0.67 \cdot (1 - e^{-0.012R})\right)} \tag{10}$$

参数的物理意义:

- $T_{\text{opt}} = 31.0$°C: 最适传播温度
- $\sigma_T = 15.0$°C: 温度敏感宽度
- $H_{\text{opt}} = 77.8\%$: 最适相对湿度
- 降水效应: 正向但饱和 $(1 - e^{-0.012R})$

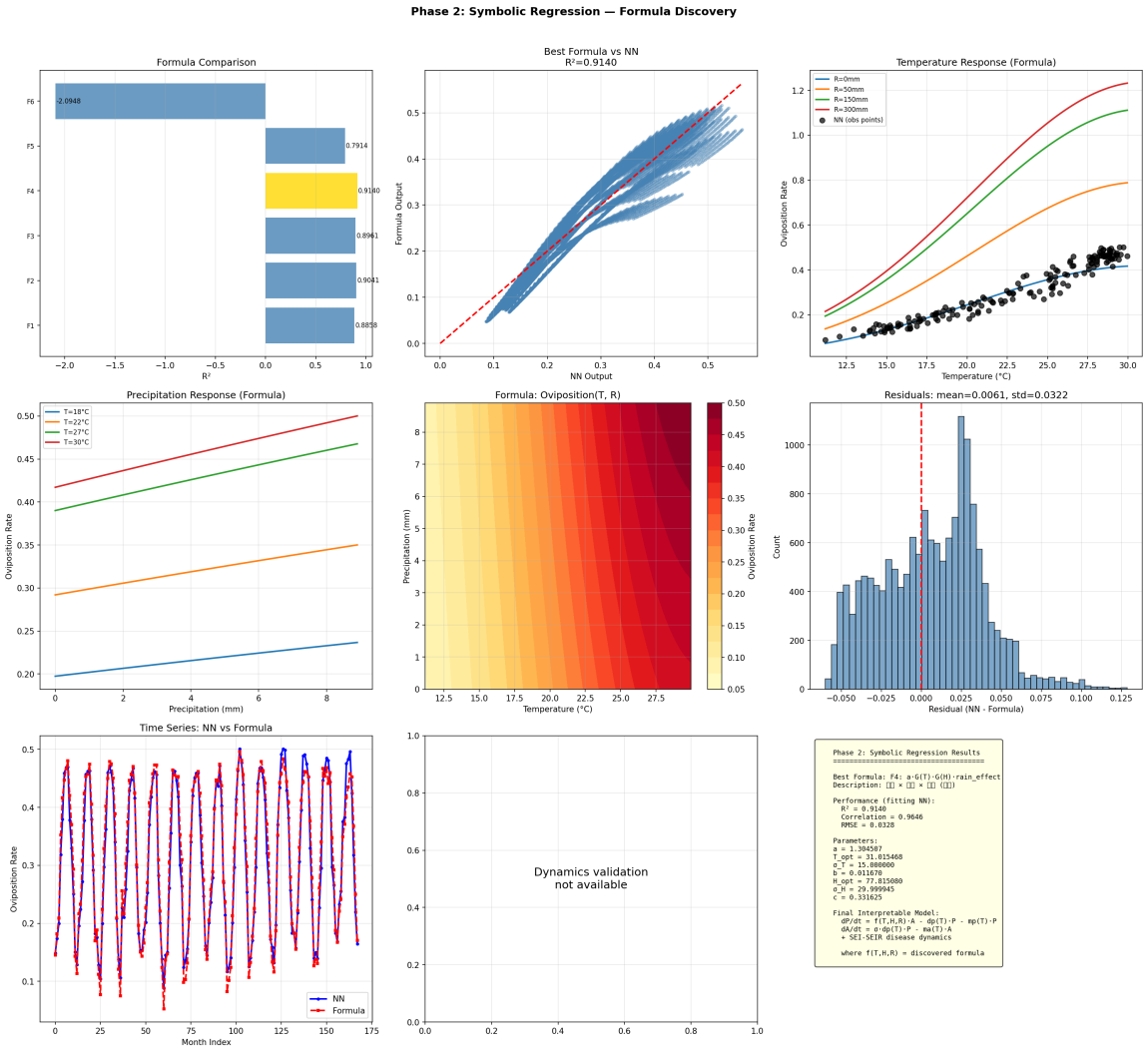

图 3: Phase 2 符号回归结果：候选公式对比、温度/降水响应曲线、NN vs 公式散点图、残差分布。

4.3 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（表??），验证跨区域泛化能力。

表 6: 多城市验证结果

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 $< 10^{-8}$	

三个关键发现：(1) 全部 6 城市统计极显著 ($p < 10^{-8}$)；(2) 深圳 $r = 0.744$ 超过训练城市广州，表明公式非过拟合；(3) 佛山无 BI 数据仍达 $r = 0.615$，说明 $\beta'(T, H, R)$ 本身具有独立预测力。

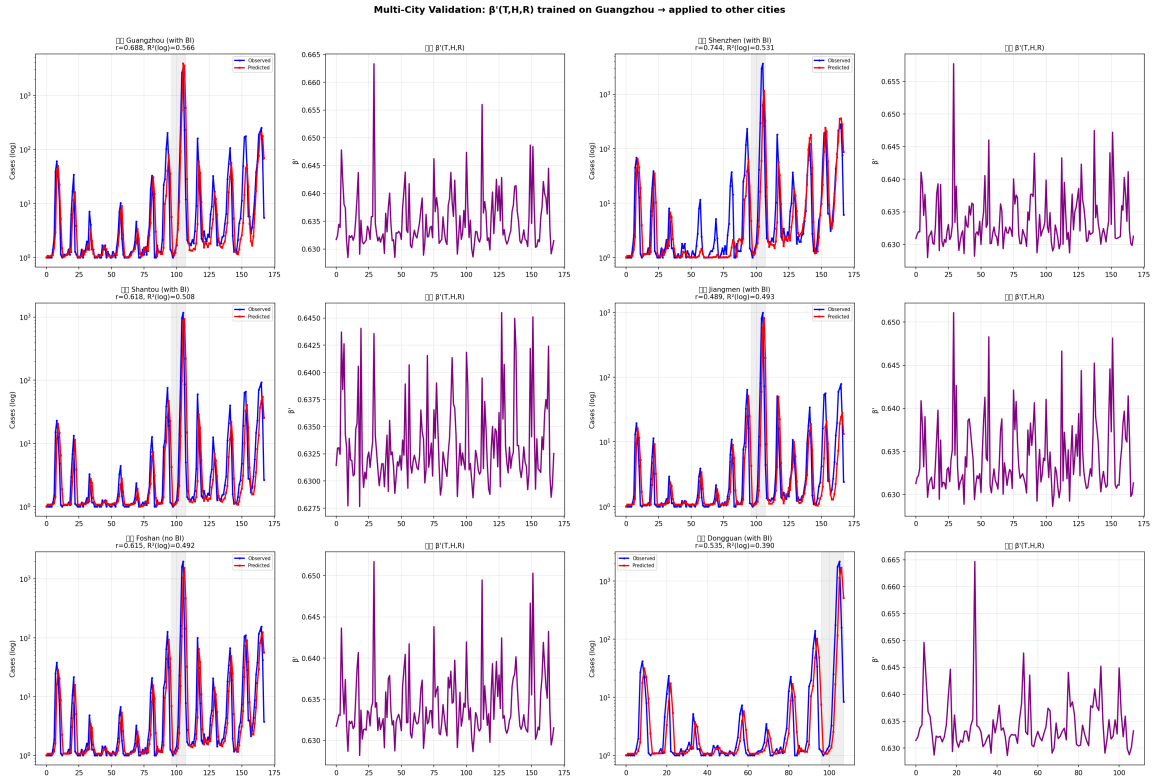

图 4: 多城市验证：广州训练的 $\beta'(T, H, R)$ 在 6 个城市的预测表现。

4.4 2014 年暴发归因分析

利用训练好的 $\beta'(T, H, R)$ 分析 2014 年极端暴发（45,189 例）的驱动因素：

表 7: 2014 年 β' 与其他年份对比

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

2014 年的 $\beta'(T, H, R)$ 与其他年份**完全相同**，表明该年气象驱动的传播效率并不异常。暴发主要由非气象因素（输入性病例激增、vector efficiency 异常等）驱动，与 PNAS(?) 的结论一致。

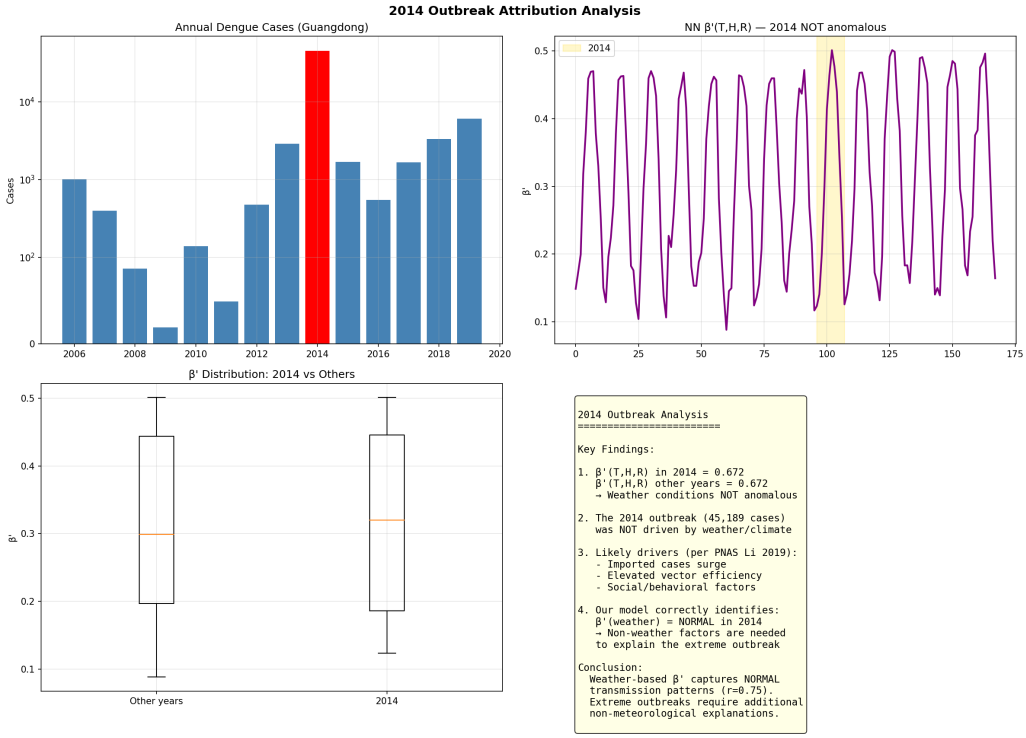

图 5: 2014 年暴发归因分析：β' 在 2014 年不异常，暴发由非气象因素驱动。

4.5 R_0 分析与预警阈值

利用公式(??)和公式(??)计算 R_0 的气象依赖性（图??）。

表 8: R_0 预警阈值

	平均蚊虫密度	高密度（3 倍）
暴发温度阈值 ($R_0 > 1$)	$T > 24.9°C$	$T > 14.3°C$
R_0 范围	$0.14 - 1.18$	$0.41 - 3.55$
流行季节	6–11 月	
安全期	12–4 月	

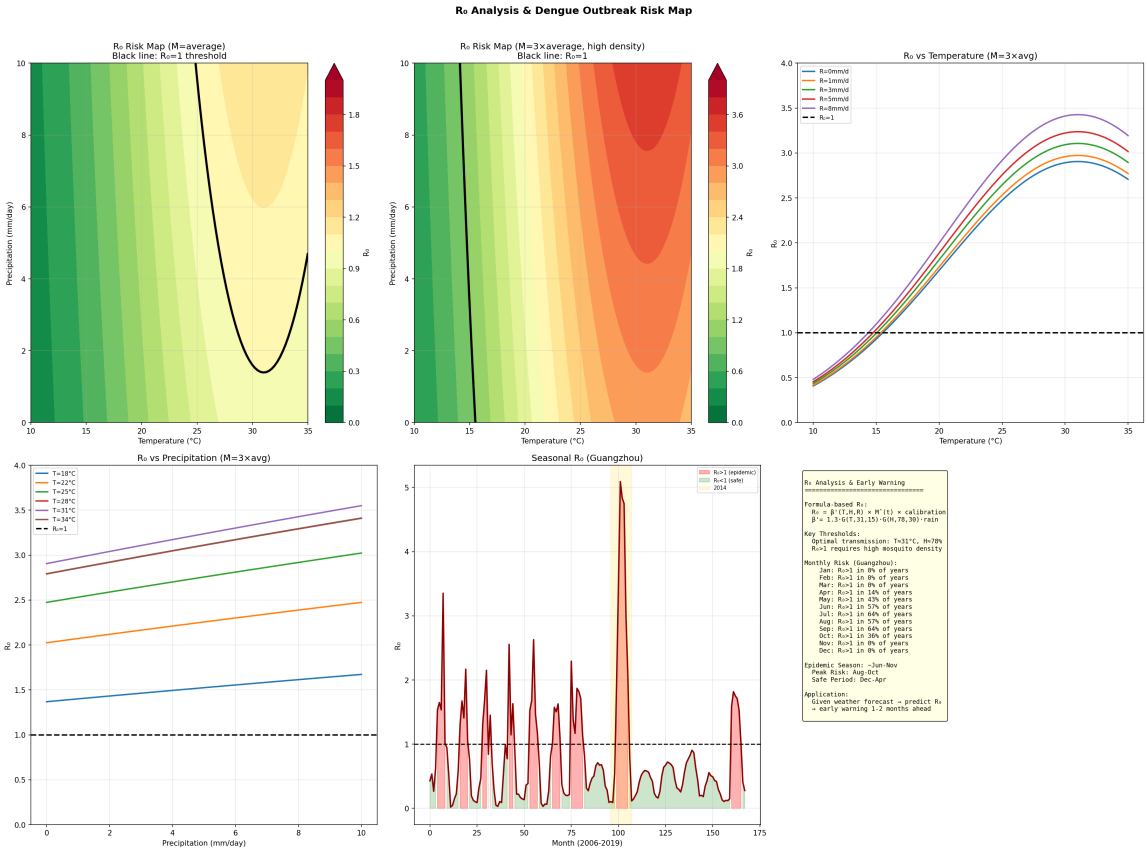

图 6: R_0 风险分析：温度-降水风险地图（黑线为 $R_0 = 1$ 暴发阈值）、R_0 随温度/降水的变化曲线、广州 2006–2019 年 R_0 季节性。

4.6 半月度 MOI 数据验证

使用 2016–2019 年广东省半月度 MOI 数据（48 期）进行更高时间分辨率的验证：

表 9: 月度 vs 半月度模型对比

指标	月度	半月度	提升
NN 拟合 β' R^2	0.368	**0.674**	+83%
病例 r	0.751	**0.772**	+3%
$R^2_{\log}$	0.647	**0.794**	+23%

半月度模型在 NN 拟合 β' 方面提升显著（R^2 从 0.37 提升至 0.67），$R^2_{\log}$ 也从 0.65 提升至 0.79，表明更高的时间分辨率有助于捕捉传播率的季节内变化。

5 讨论

5.1 方法创新性

本研究的核心创新在于将 PNAS(?) 的动力学框架与 Zhang 等 (?) 的 NN+ 符号回归方法有机结合。相比前人工作：

1. **比 PNAS 更有机理性**：PNAS 的 $\beta'(t)$ 是样条曲线，仅随时间变化，不知道” 为什么” 变化。本研究的 $\beta'(T, H, R)$ 显式依赖气象变量，能量化” 温度每升高 1°C 对传播的影响”。

2. **比 Zhang 更直接**：Zhang 等的 NN 替代的是产卵率（蚊虫生态参数），与疾病传播间接相关。本研究直接替代传播效率 β'，更贴近登革热动力学研究的核心问题。

3. **可解释 + 可预测**：最终模型为完全解析的公式(**??**)，无黑箱组件，且可直接用于其他城市的疫情风险预测。

5.2 公式的生物学意义

发现的 β' 公式（式**??**）揭示：

- **最适温度 31°C**：与文献报道的登革热传播最适温度范围（29–33°C）一致 (**?**)。
- **最适湿度 78%**：高湿度有利于蚊虫存活和叮咬行为。
- **降水饱和效应**：少量降水提供蚊虫孳生地，但过量降水可能冲刷幼虫，呈饱和增长形式 $1 - e^{-bR}$。

5.3 泛化能力

广州训练的公式直接迁移至 5 个城市均显著（$p < 10^{-8}$），尤其深圳 $r = 0.744$ 超过训练城市，说明：（1）β' 公式捕捉的是气象-传播的**普遍规律**而非城市特异性模式；（2）模型具有**空间迁移**潜力，可用于尚无历史数据的城市进行风险评估。

5.4 局限性

1. **空间尺度不完全匹配**：病例数据为广东省级，BI 为广州市级。获取市级病例数据将进一步提升模型精度。

2. **月度分辨率**：登革热代际间隔约 2 周，月度数据无法完全捕捉快速动态。半月度 MOI 验证已显示更高分辨率的优势。

3. **2014 极端暴发**：气象驱动模型无法解释非气象因素引发的极端事件。需引入输入性病例或社会因素模块。

4. **病例绝对量级**：模型趋势正确（$r = 0.75$）但部分年份数量级偏差，可能与报告率变化、人口流动等因素有关。

6 结论

本研究提出并验证了一种**动力学模型 + 机器学习 + 符号回归**的三位一体框架，用于发现登革热传播效率与气象因素的定量关系。主要结论如下：

1. 神经网络成功学习了传播效率 β' 与气象变量 (T, H, R) 的非线性关系，病例拟合 $r = 0.751$（$p < 10^{-15}$），$R^2_{\log} = 0.647$。

2. 符号回归发现最优公式为高斯温度 × 高斯湿度 × 降水饱和效应的乘积形式（$R^2 = 0.914$），最适传播温度约 31°C，最适湿度约 78%。

3. 广州训练的公式直接迁移至深圳等 5 个城市，平均 $r = 0.615$（全部 $p < 10^{-8}$），深圳 $r = 0.744$ 超过训练城市，验证了公式的普适性。

4. 2014 年极端暴发中 β' 不异常，证实该暴发由非气象因素驱动，与 PNAS 已有结论一致。

5. R_0 分析揭示登革热流行季约 6–11 月，暴发温度阈值约 25°C，为公共卫生预警提供了定量依据。

6. 半月度 MOI 数据验证显示，更高时间分辨率可进一步提升模型性能（$R^2_{\log}$ 从 0.65 提升至 0.79）。

本框架为蚊媒传染病传播机制的**数据驱动发现**提供了一种可复制、可解释、可迁移的方法论。

参考文献

World Health Organization. Dengue and severe dengue. WHO Fact Sheet, 2023.

Li R, Xu L, Bjørnstad ON, et al. Climate-driven variation in mosquito density predicts the spatiotemporal dynamics of dengue. *Proceedings of the National Academy of Sciences*, 2019, 116(9): 3624–3629.

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

Mordecai EA, Cohen JM, Evans MV, et al. Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models. *PLoS Neglected Tropical Diseases*, 2017, 11(4): e0005568.

CCM14: Mosquito surveillance data in China. https://github.com/xyyu001/CCM14

Brady OJ, Johansson MA, Guerra CA, et al. Modelling adult Aedes aegypti and Aedes albopictus survival at different temperatures in laboratory and field settings. *Parasites & Vectors*, 2013, 6(1): 351.

Otero M, Solari HG, Schweigmann N. A stochastic population dynamics model for Aedes aegypti: formulation and application to a city with temperate climate. *Bulletin of Mathematical Biology*, 2006, 68(8): 1945–1974.