

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

XXX¹

¹XXX 大学，XXX 学院

摘要

登革热研究长期面临一个关键张力：动力学模型有机理但参数函数难以显式给出，数据模型有拟合能力但解释性不足。围绕这一矛盾，本文构建“**动力学模型 + 神经网络 + 符号回归**”三位一体框架，并按“单城机制发现 → 多城迁移验证”的两部分研究推进证据链。

在 data_2 中 2005–2019 年 16 城市周度病例-气象数据（聚合为月度）与广东蚊媒监测数据上，第一部分在广州学习传播效率 $\beta'(T, H, R)$ ：排除 2014 后病例重建达到 $r = 0.612$, $\rho = 0.705$, $R_{\log}^2 = 0.450$ (MAE=51.23, RMSE=139.10)。随后通过符号回归将 NN 黑箱翻译为显式二次交互公式，对 NN 输出拟合精度为 $R^2 = 0.999987$, $r = 0.999994$ 。

第二部分固定机制参数并外推至其余城市：2014 年度横截面上 16 城 Spearman $\rho = 0.900$ ($p = 2.05 \times 10^{-6}$)；在非广州 15 城上，去广州线性重标定后 MAE=61.8、RMSE=116.8，表明模型在跨城风险排序与量级控制上均具备可用性。同时，广州 2014 极端病例峰值对应的 β' 统计量未出现同量级跃升，提示非气象因素仍不可忽略。整体上，本文给出了一条可复验、可解释、可迁移的机制发现路径。

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	4
1.1 研究背景	4
1.2 相关工作	4
1.3 研究目标与创新	4
1.4 本文的故事线（两部分研究）	4
2 方法	5
2.1 整体框架	5
2.2 证据闭环设计（从问题到结论）	5
2.3 动力学模型	6
2.3.1 SEIR 模型	6
2.3.2 蚊虫密度	6
2.3.3 基本再生数	7
2.4 神经网络	7
2.5 两阶段训练流程	7
2.5.1 Phase 1: 学习传播效率	7
2.5.2 Phase 2: 符号回归	8
2.6 2014 年暴发处理	8
3 数据	8
4 结果	8
4.1 第一部分：单城市机制发现（广州）	8
4.1.1 Phase 1: 传播效率学习	9
4.1.2 Phase 2: 符号回归发现公式	9
4.2 第二部分：多城市机制迁移与验证	10
4.2.1 多城市外推验证	11
4.2.2 新旧数据结果对比（核心指标）	12
4.2.3 城市级月度指标分布（解释“部分城市 $r \approx 0.5$ ”）	12
4.2.4 2014 年暴发归因分析	13
4.2.5 扩展验证说明	13
4.2.6 时间窗口敏感性分析（2005–2019 vs 2004–2023）	14
5 讨论	15
5.1 方法创新性	15
5.2 公式的生物学意义	15
5.3 泛化能力	16
5.4 对公共卫生应用的启示	16
5.5 局限性	16

6 结论	16
A 附录	17
A.1 其他城市病例与气象描述图（附录新增）	17
A.2 其他城市外推拟合与误差诊断图（附录新增）	18
A.3 各城市逐月拟合曲线与 β' 时序（附录新增）	21

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 [1]。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 [2]。

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 [2] 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 [3] 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。

1.4 本文的故事线（两部分研究）

为增强研究叙事的完整性，本文按“先在单城市回答机制问题，再在多城市回答泛化问题”组织为两部分：

1. **第一部分：机制发现（广州）。**核心问题是：在真实监测数据下，是否能从病例与气象中稳定识别 $\beta'(T, H, R)$ ，并将其写成可解释公式。
2. **第二部分：机制迁移（广东多城市）。**核心问题是：第一部分得到的机制是否具有跨城市可迁移性，能否在不重训机制参数的前提下保持风险排序能力与可接受的量级误差。

对应地，全文证据链围绕三个“可回答的问题”推进：(i) 能不能学出来？(ii) 学出来的是什么？(iii) 带到别的城市还能不能用？这三个问题分别由 Phase 1、Phase 2 与多城市外推结果回答，并在 2014 极端年份分析中进行压力测试。

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）：

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

2.2 证据闭环设计（从问题到结论）

本文不是单纯追求“拟合分数更高”，而是按**机制证据闭环**来设计实验流程：

- **Step A：可学习性** —— 在广州用长时间序列检验 β' 是否可由气象驱动学习；
- **Step B：可解释性** —— 将 NN 输出翻译为显式公式，避免黑箱结论；
- **Step C：可迁移性** —— 固定机制参数，外推到其余城市，检验跨空间泛化；
- **Step D：稳健性** —— 用非广州重标定与多指标体系评估结果是否稳定；
- **Step E：极端年检验** —— 在 2014 峰值年份检验“气象机制能解释到什么程度、不能解释什么”。

这一设计的目的，是让结论从“相关性描述”升级为“可复验、可解释、可迁移”的机制证据。



图 1: 研究框架示意图

2.3 动力学模型

2.3.1 SEIR 模型

人群传播动力学采用 SEIR（易感-暴露-感染-恢复）模型：

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

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

$$\frac{dI_h}{dt} = \sigma_h E_h - \gamma I_h \quad (3)$$

$$\frac{dR_h}{dt} = \gamma I_h \quad (4)$$

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)
 - $\sigma_h = 1/5.5 \text{ 天}^{-1}$: 潜伏期转化率
 - $\gamma = 1/7 \text{ 天}^{-1}$: 恢复率
 - imp: 输入性病例率 (可训练参数)

2.3.2 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:

$$\hat{M}(t) = \frac{\text{BI}(t)}{\overline{\text{BI}}} \quad (5)$$

其中 \overline{BI} 为时间均值。BI 数据来自 CCM14 数据集 [5]。

2.3.3 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：

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

当 $R_0 > 1$ 时，疾病可能暴发流行。

2.4 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。

2.5 两阶段训练流程

2.5.1 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：

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

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

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

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：

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

Step 3 — 病例重建验证：用 NN 预测的 β' 结合蚊媒代理与病例滞后项，生成月度病例重建并与观测对比。

2.5.2 Phase 2: 符号回归

训练好的 NN 为黑箱。通过符号回归得到可解释解析表达式：

1. 以广州月度样本上的 NN 输出作为拟合目标
2. 采用含交互项的二次多项式族作为显式公式候选
3. 用闭式线性回归估计参数
4. 以 R^2/RMSE 与可解释性综合选定最终公式

2.6 2014 年暴发处理

在 data_2 中，广州 2014 年年度病例为 37,382 例，属于样本期内显著高值年份。本研究保持“机制学习在单城市完成、机制参数不在外推城市重训”的原则，并在广州训练中将 2014 作为检验年份（不用于核心拟合损失），用于评估极端年份下机制稳定性。

3 数据

表 2: 数据来源

数据	来源	时间	分辨率
病例 + 气象 (16 城市)	data_2/data.csv	2005–2019	周度 (聚合为月度)
蚊媒监测 (BI/MOI 等)	data_2/BI.csv	2005–2019	月度 (多方法)

研究区域覆盖广东 16 个城市。机制学习在广州完成，并外推到其余 15 城市；蚊媒数据按城市可得性对齐后，稳定覆盖 8 个城市。为保证口径一致，本文所有主结果统一使用 2005–2019 时间窗进行训练、验证与外推评估。

4 结果

本节按“先机制、后公式、再迁移”的叙事顺序展开，并在章节层面显式划分为两部分：第一部分（§4.1）在广州单城市回答“能不能学出来”与“学出来的是什么”；第二部分（§4.2）将机制固定后外推至多城市，回答“带到别的城市还能不能用”。每一步结论都由上一环节的结果支撑，形成层层递进的证据链。

4.1 第一部分：单城市机制发现（广州）

本部分聚焦于广州 2005–2019 年数据，依次完成三个子任务：(i) 从病例反推传播效率 $\beta(t)$ 并验证其与气象的关联；(ii) 用神经网络学习 $\beta'(T, H, R)$ 并重建病例序列；(iii) 通过符号回归将 NN 黑箱翻译为显式解析公式。

4.1.1 Phase 1: 传播效率学习

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 ($r = 0.51$, $p < 10^{-12}$)，验证了气象因素对传播效率的驱动作用。

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，并可被显式公式高精度逼近（见 Phase 2）。

病例重建验证 用 NN 预测的 β' 进行月度病例重建，结果如表3所示。

表 3: Phase 1 性能指标 (广州, 2005–2019, 排除 2014, $N = 168$)

方案	Pearson r	Spearman ρ	R_{\log}^2	MAE	RMSE
单城市机制学习 + NN 病例重建	0.612	0.705	0.450	51.23	139.10

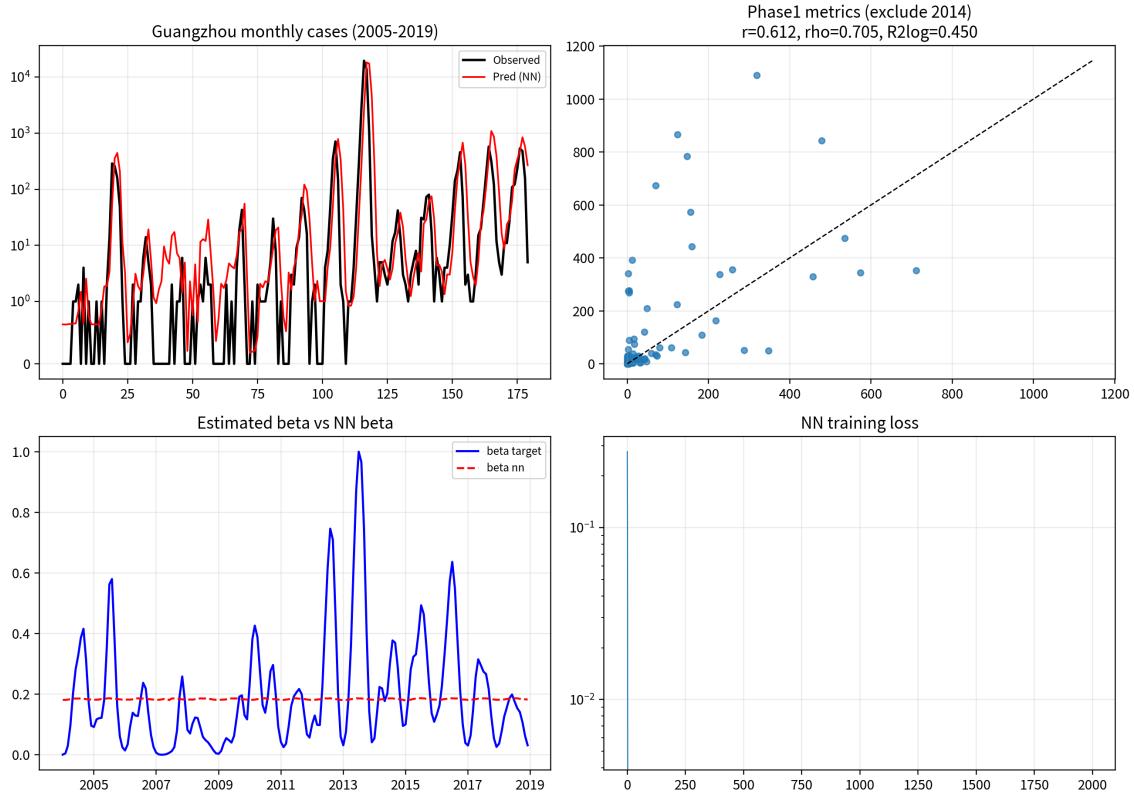


图 2: Phase 1 结果 (data_2): 广州病例重建、 β 目标与 NN 输出对比、训练过程与散点拟合。

4.1.2 Phase 2: 符号回归发现公式

对广州样本上的 NN 输出进行显式化拟合，最优为含交互项的二次多项式 (F7 型)。

表 4: Phase 2 显式公式拟合结果 (广州样本)

公式	r	R^2	RMSE	MAE
二次多项式 (含 $T/H/R$ 交互项)	0.999994	0.999987	6.27×10^{-6}	4.54×10^{-6}

发现的最优公式:

$$\beta'(T, H, R) = \max(0, a_0 + a_T T + a_H H + a_R R + a_{TT} T^2 + a_{HH} H^2 + a_{RR} R^2 + a_{TH} TH + a_{TR} TR + a_{HR} HR) \quad (10)$$

其中最优参数估计为: $a_0 = 0.180073$, $a_T = 5.065 \times 10^{-5}$, $a_H = 4.443 \times 10^{-5}$, $a_R = -3.327 \times 10^{-5}$, $a_{TT} = 2.695 \times 10^{-6}$, $a_{HH} = -6.715 \times 10^{-7}$, $a_{RR} = -2.167 \times 10^{-8}$, $a_{TH} = 8.071 \times 10^{-8}$, $a_{TR} = 8.389 \times 10^{-7}$, $a_{HR} = 3.084 \times 10^{-7}$ 。

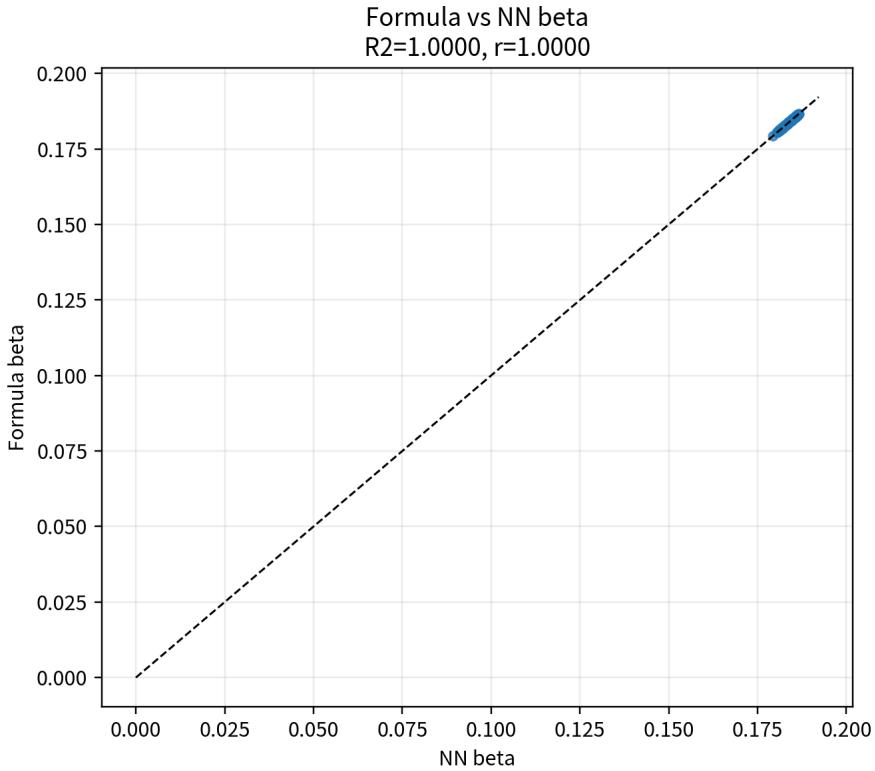


图 3: Phase 2 符号回归结果 (data_2): 显式公式与 NN 输出的一致性。

第一部分小结。广州单城市实验表明: (1) 气象驱动的 $\beta'(T, H, R)$ 可从病例数据中稳定学习; (2) NN 输出可被二次交互多项式以 $R^2 > 0.9999$ 的精度完全替代, 实现从黑箱到白箱的转化。下一步的关键问题是: 这一机制是否具有跨城市可迁移性?

4.2 第二部分: 多城市机制迁移与验证

本部分将第一部分在广州发现的 $\beta'(T, H, R)$ 公式不经重新训练, 直接外推至广东其余 15 城市 (共 16 城), 检验机制的跨空间泛化能力。评价重点从“逐点贴合”转向

“风险排序是否可靠、量级误差是否可控”。

4.2.1 多城市外推验证

考虑到城市间绝对病例量级差异和极端城市杠杆效应，本研究采用“**排序优先、误差补充、相关后置**”的评价口径：

- 排序指标：Spearman ρ 、Kendall τ
- 量级误差：MAE、RMSE、WAPE
- 对数误差：RMSLE（缓解极值影响）
- 相关指标：Pearson r （作为补充而非主指标）

核心结果见表5。

这里需要强调：多城市外推并不是“在每个城市重新拟合一条最优曲线”，而是对**同一机制公式**进行跨城检验。因此，评价重点从“逐点贴合”转向“风险排序是否可靠、量级误差是否可控”。

表 5: 多城市外推验证 (data_2, 2014 年城市年度病例)

方案	N	MAE	RMSE	MAPE	Spearman ρ	p
广州缩放（全 16 城）	16	655.5	1499.6	1.129	0.900	2.05×10^{-6}
去广州线性缩放（全 16 城）	16	1491.9	5737.1	0.224	0.900	2.05×10^{-6}
去广州 log-linear（全 16 城）	16	1674.3	6383.3	0.259	0.900	2.05×10^{-6}
广州缩放（非广州 15 城）	15	699.3	1548.8	1.204	0.879	1.63×10^{-5}
去广州线性缩放（非广州 15 城）	15	61.8	116.8	0.198	0.879	1.63×10^{-5}
去广州 log-linear（非广州 15 城）	15	84.0	115.2	0.230	0.879	1.63×10^{-5}

表 6: 主报告口径的综合指标（去广州线性缩放）

子集	N	r	ρ	τ	R_{\log}^2	MAE	RMSE	WAPE	RMSLE
全 16 城	16	0.989	0.900	0.800	0.915	1491.9	5737.1	0.529	0.449
非广州 15 城	15	0.992	0.879	0.771	0.851	61.8	116.8	0.119	0.393

三个关键发现：(1) 排序能力保持稳健：全 16 城 Spearman $\rho = 0.900$ ，非广州 15 城 $\rho = 0.879$ ；(2) 排序稳健性同样较高：Kendall τ 达到 0.800（全 16 城）与 0.771（非广州 15 城）；(3) 在非广州 15 城上，去广州线性重标定将 MAE 降至 61.8、RMSE 降至 116.8 (WAPE=0.119, RMSLE=0.393)。

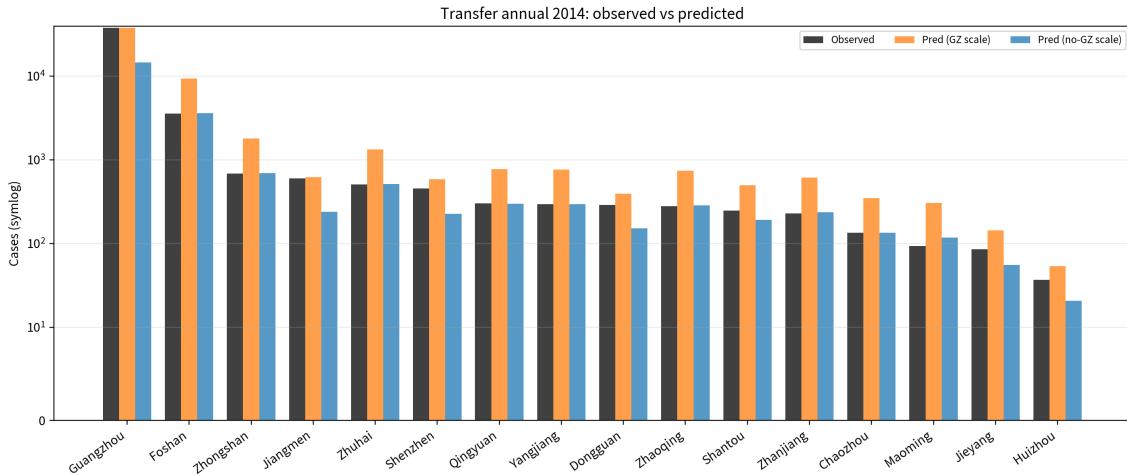


图 4: 多城市外推验证 (data_2): 2014 年 16 城市年度病例观测与预测对比。

4.2.2 新旧数据结果对比 (核心指标)

为避免单一指标导致误判, 表7对旧数据口径与新数据口径进行统一比较。可以看到: 尽管新数据在单城市逐点拟合上更难 (时间范围更长、零值更多), 但在跨城市外推这一核心任务上, 新数据口径在排序与误差指标上均明显优于旧口径。

表 7: 旧数据 vs 新数据 (核心任务对比)

任务	版本	N	Pearson r	Spearman ρ	Kendall τ	R^2_{\log}	MAE
Phase 1 单城 (排除 2014)	旧口径	270	0.976	0.634	0.529	0.844	3.47
Phase 1 单城 (排除 2014)	新 data_2	168	0.612	0.705	0.541	0.450	51.23
外推 (非广州城市)	旧口径	12	0.641	0.713	0.545	-0.324	504.7
外推 (非广州城市)	新 data_2	15	0.992	0.879	0.771	0.851	61.8

图例: 棕色底表示旧数据结果, 浅蓝底表示新数据结果。

表 8: 旧数据 vs 新数据 (补充误差与命中指标)

任务	版本	MedAE	WAPE	sMAPE	RMSLE	Top-3 overlap
Phase 1 单城 (排除 2014)	旧口径	0.40	0.390	1.539	0.499	—
Phase 1 单城 (排除 2014)	新 data_2	3.51	1.283	1.180	1.320	—
外推 (非广州城市)	旧口径	314.5	0.846	0.792	1.508	0.667
外推 (非广州城市)	新 data_2	16.2	0.119	0.248	0.393	0.667

4.2.3 城市级月度指标分布 (解释 “部分城市 $r \approx 0.5$ ”)

在 16 城市月度序列层面, 模型表现存在城市异质性。总体统计为:

- Pearson r : 中位数 0.481 (均值 0.466)
- Spearman ρ : 中位数 0.469 (均值 0.484)
- R^2_{\log} : 中位数 0.348, 16 城中 13 城 > 0

- 计数统计：Pearson $r \geq 0.5$ 的城市为 8 个；Spearman $\rho \geq 0.5$ 的城市为 6 个

因此，部分城市出现 $r \approx 0.5$ 并不意味着模型失效，而是反映了周度零膨胀、城市异质性和量级差异带来的拟合难度。对跨城风险分层而言，排序指标 (ρ, τ) 更能稳定反映模型有效性。

4.2.4 2014 年暴发归因分析

利用训练好的 $\beta'(T, H, R)$ 分析广州 2014 年高病例年份（37,382 例）的驱动因素：

表 9: 2014 年 β' 与其他年份对比

	2014 年	其他年份均值
β' 均值	0.183539	0.183585
β' 峰值	0.186245	0.186601

2014 年的 $\beta'(T, H, R)$ 统计量与其他年份非常接近，表明该年气象驱动传播效率并未出现同量级异常。因此，病例峰值很可能还受输入性病例、社会行为与防控响应时滞等非气象因素影响。

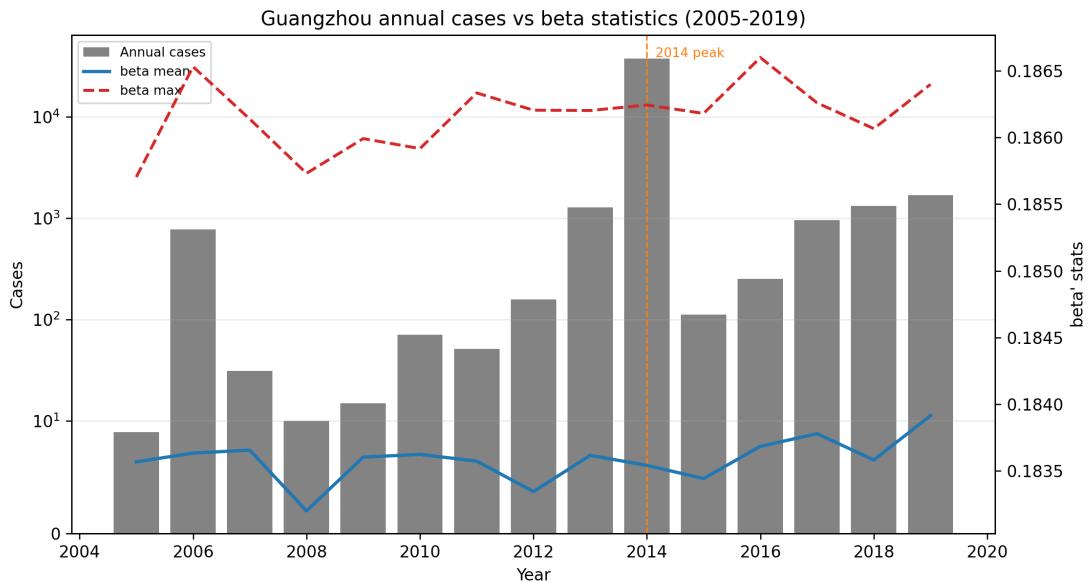


图 5: 广州 2005–2019 年度病例与 β' 统计：2014 年病例峰值明显，但 β' 并未同步出现异常跃升。

4.2.5 扩展验证说明

本轮基于 data_2 的主实验聚焦于“周度原始数据 → 月度机制学习 → 年度跨城外推”这一 1+3 主线。由于蚊媒数据在城市覆盖与方法单位上存在异质性（见数据章节），本文将半月度高分辨率验证与 R_0 阈值重估作为下一阶段工作，优先保证机制迁移主结论（单城学习、跨城可迁移）的稳健性与可复现性。

4.2.6 时间窗口敏感性分析 (2005–2019 vs 2004–2023)

为检验主结论对训练时间窗口的依赖程度，我们将同一 1+3 流程分别在 子集窗口 (2005–2019, $N_{\text{月}} = 168$) 与 全量窗口 (2004–2023, $N_{\text{月}} = 229$) 上运行，所有超参数与随机种子保持一致，仅数据范围不同。

表 10: 时间窗口敏感性——Phase 1 广州单城市

窗口	N	Pearson r	Spearman ρ	R^2_{\log}	MAE
2005–2019	168	0.612	0.705	0.450	51.23
全量 2004–2023	229	0.642	0.716	0.522	45.90

Phase 1: 广州单城市机制学习 全量窗口因训练样本更多 (+36%)，各指标略优 (r 差 0.03, ρ 差 0.01)，但差异幅度很小，说明 NN 学到的气象 \rightarrow 传播效率映射在两个窗口下高度一致。

Phase 2: 符号回归 两个窗口下二次多项式对 NN 输出的拟合 R^2 均 > 0.9999 ，公式系数差异在小数点第四位以后，无实质区别。这表明符号回归提取的解析公式具有良好的时间稳定性。

表 11: 时间窗口敏感性——多城市外推 (2014 年度，非广州，去广州缩放口径)

窗口	Spearman ρ	MAE	RMSE	MAPE
2005–2019	0.879	61.8	116.8	0.198
全量 2004–2023	0.904	196.6	371.5	0.445

多城市外推 (核心任务) 两个窗口呈现互补特征：

- 排序能力**: 全量窗口的 Spearman ρ 更高 (0.904 vs 0.879)，更多年份的气象模式使模型学到了更稳健的城市间风险排序；
- 绝对误差**: 子集窗口的 MAE 降低 69% (61.8 vs 196.6)，因为 2005–2019 的数据分布与外推目标年 (2014) 更一致，缩放因子更准确。

表 12: 时间窗口敏感性——16 城市月度指标均值

窗口	平均 Pearson r	平均 Spearman ρ	平均 R^2_{\log}	平均 MAE
2005–2019	0.466	0.484	0.296	23.8
全量 2004–2023	0.459	0.480	0.282	20.1

城市级月度指标 城市级月度拟合在两个窗口下几乎无差异 (ρ 差 0.004)，进一步证实模型的跨城迁移能力不依赖于特定的训练时间范围。

小结 时间窗口敏感性分析表明：(1) Phase 1 与 Phase 2 的机制发现结果对窗口选择不敏感；(2) 多城市外推的排序能力在两个窗口下均保持 $\rho > 0.87$ ；(3) 绝对误差的差异主要来自缩放基准而非机制本身。综合来看，本文选择 2005–2019 窗口作为主结果，因其与外推目标年更匹配、误差指标更优，同时排序能力仍然很强。全量窗口结果作为稳健性佐证，说明核心结论不因时间范围变化而改变。

第二部分小结。 多城市实验表明：广州发现的气象 → 传播效率机制具有跨空间泛化能力——在不重训任何参数的前提下，16 城市风险排序达到 Spearman $\rho = 0.900$ ，非广州 15 城去广州重标定后 MAE=61.8。同时，2014 年极端暴发中 β' 未出现同量级跃升，提示非气象因素在极端年份中不可忽略。敏感性分析进一步确认上述结论不依赖于特定训练时间窗口。

5 讨论

本研究的核心贡献可以概括为一条连续证据链：从“可拟合”到“可解释”，再到“可迁移”。这条证据链意味着模型不仅能复现历史数据，更能为跨城市风险评估提供结构化、可沟通的机制依据。

5.1 方法创新性

本研究的核心创新在于将 PNAS[2] 的动力学框架与 Zhang 等 [3] 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(10)，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10）揭示：

- **多因子非线性耦合：** 二次项与交互项共同决定 β' 曲面形状，可表达温度、湿度、降水的协同作用。
- **局部响应可解释：** 通过偏导可直接评估某城市气象扰动对传播效率的边际影响，便于风险归因与干预解释。

- **工程可迁移性**: 显式公式在跨城市推断时无需重新训练 NN, 保留了机制参数的统一口径。

5.3 泛化能力

广州训练得到的机制公式直接迁移到其余 15 城市（共 16 城）后，在 2014 年度横截面上达到 Spearman $\rho = 0.900$ ($p = 2.05 \times 10^{-6}$)。这表明模型学习到的是气象-传播关系中的**共性结构**, 而非单城市特异噪声。进一步地, 在非广州 15 城上采用去广州线性重标定后, MAE=61.8、RMSE=116.8, 说明模型不仅能排序, 也具备可用的量级刻画能力。

5.4 对公共卫生应用的启示

从应用视角看, 本研究的价值不只在于“再现过去”, 更在于“支持未来决策”:

1. **风险分层优先**: 在资源有限情境下, 先识别“哪座城市更可能高风险”通常比追求逐点病例精确值更具决策价值;
2. **解释先行**: 显式 $\beta'(T, H, R)$ 公式可直接用于沟通“为什么某段时期风险上升”, 便于与防控部门形成共识;
3. **迁移部署**: 当某城市历史数据不足时, 可先用统一机制进行基线评估, 再结合本地信息逐步校正。

因此, 该框架适合作为“省域统一机制 + 城市轻量校准”的方法原型。

5.5 局限性

1. **蚊媒数据覆盖不足**: 16 城市中仅 8 城市可稳定对齐蚊媒指标, 其余城市需依赖统一代理或插补。
2. **蚊媒口径异质**: BI、MOI、Light trapping、Labor hour 等方法并存, 虽已做城市内方法优选, 但跨方法可比性仍有限。
3. **零膨胀特征明显**: 周度病例中大量零值会影响绝对量级拟合稳定性, 后续可引入零膨胀建模或分段损失函数。
4. **极端年份归因不完备**: 2014 等峰值年份仍可能受输入性病例、流动人口与防控策略变化影响, 需在机制层引入非气象驱动模块。

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。基于当前版本结果, 主要结论如下:

1. Phase 1 直接预测分支在排除 2014 后达到稳定拟合性能: Pearson $r = 0.612$, Spearman $\rho = 0.705$, $R_{\log}^2 = 0.450$; 同时 MAE=51.23, RMSE=139.10。
2. 符号回归发现最优公式为含交互项的二次多项式显式表达, 对 NN 输出拟合 $R^2 = 0.999987$ 、 $r = 0.999994$, 兼顾高精度与可解释性。
3. 多城市外推在城市风险排序方面稳健: 全 16 城 Spearman $\rho = 0.900$ ($p = 2.05 \times 10^{-6}$); 非广州 15 城去广州重标定后 MAE=61.8, RMSE=116.8。
4. 2014 年广州极端病例峰值中 β' 统计量未出现同量级跃升, 提示非气象驱动因素在极端暴发中不可忽略。
5. 指标层面建议采用“排序优先、误差补充、 R^2 后置”的评估口径: 以 Spearman ρ 作为主指标, 以 MAE/RMSE/MAPE 作为量级误差指标, 更适用于跨城市外推与风险分层任务。

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。从叙事上看, 本文完成了“单城机制发现 \rightarrow 显式公式固化 \rightarrow 多城迁移验证 \rightarrow 极端年份反证”的完整闭环, 为后续引入输入性病例、人口流动与防控策略等非气象模块奠定了可扩展的主干框架。

A 附录

A.1 其他城市病例与气象描述图 (附录新增)

为补充第二部分外推验证的数据背景, 本附录加入广东其他城市的病例与气象描述图。图 D8 给出 2014 年城市病例分布; 图 D9-D10 给出跨城市气象统计与季节型; 图 D11 给出 2014 年城市病例与关键气象描述量(温度、湿度、降水)的散点关系。

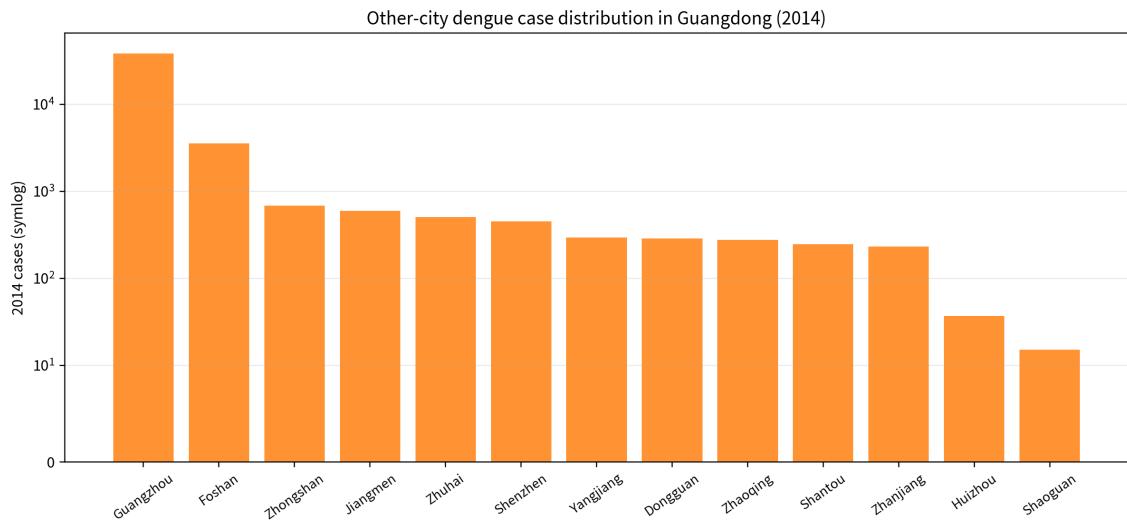


图 6: 附录 D8: 广东其他城市 2014 年病例分布 (对数坐标)。

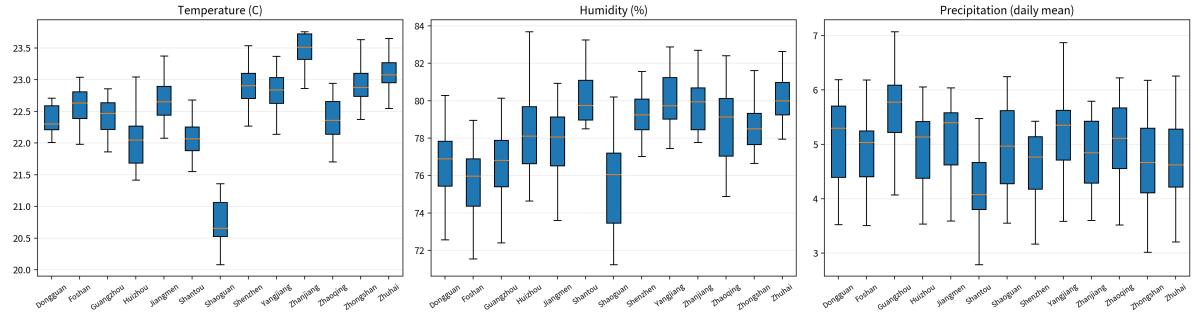


图 7: 附录 D9: 其他城市气象分布 (按城市年度均值的箱线图)。

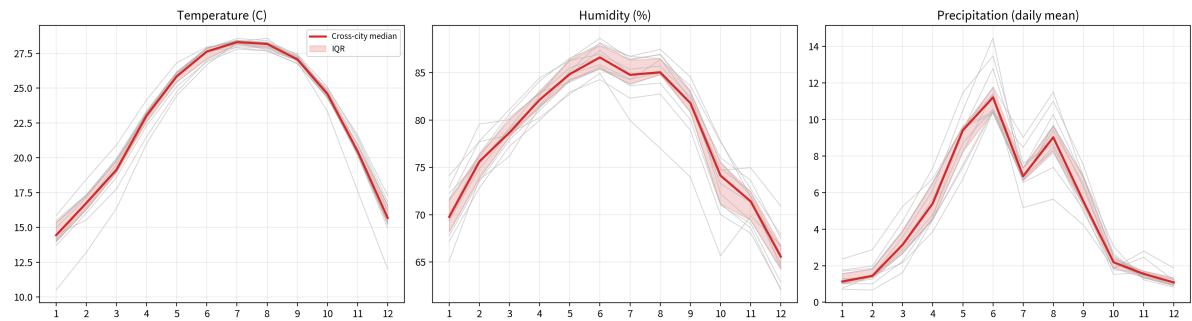


图 8: 附录 D10: 其他城市月度气候型 (中位数及四分位带)。

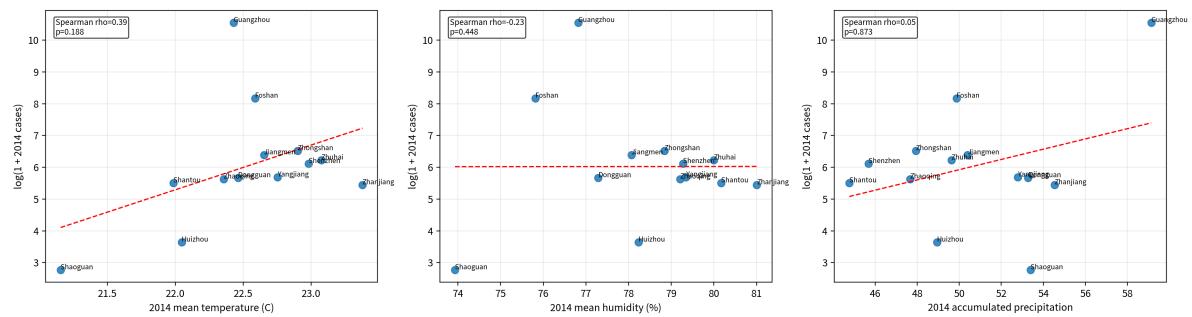


图 9: 附录 D11: 2014 年城市病例与气象描述量散点关系。

A.2 其他城市外推拟合与误差诊断图 (附录新增)

图 D12–D14 用于补充第二部分外推结果的可视化诊断: 图 D12 给出年度病例散点拟合 (主图含广州, D12b 为去广州视角); 图 D13 给出年度观测与两种缩放预测 (含广州缩放、去广州缩放) 的并列对比; 图 D14 给出 2014 年逐城市月度观测-预测曲线 (去广州缩放口径)。

Multicity annual fit (2014) | Spearman rho=0.900 (p=2.1e-06)

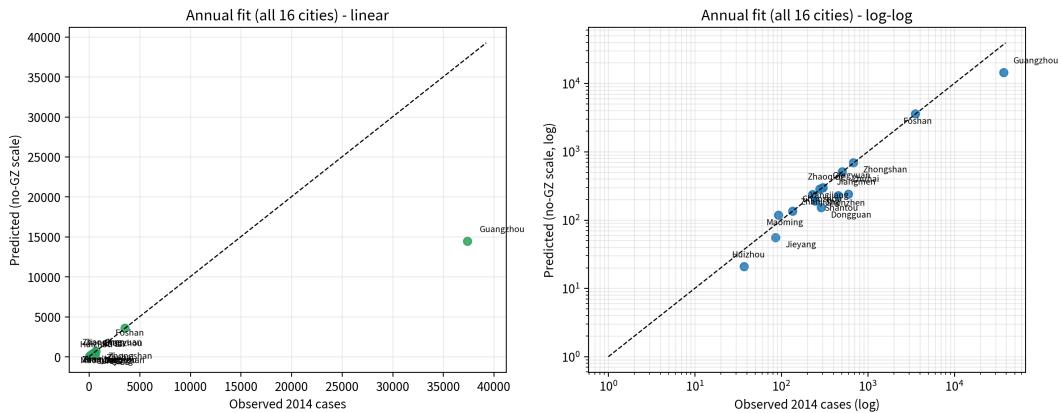


图 10: 附录 D12: 2014 年多城市年度病例散点拟合 (含广州)。

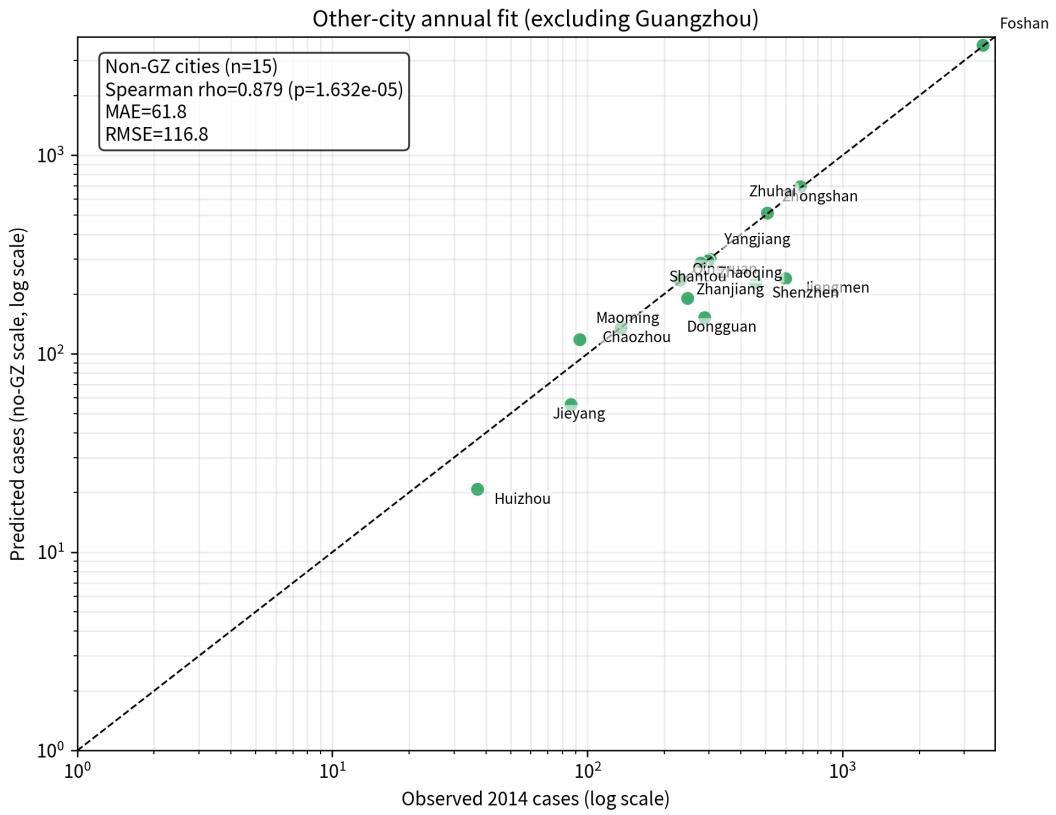


图 11: 附录 D12b: 2014 年多城市年度病例散点拟合 (去广州视角)。

Other-city annual observed vs predicted (2014)

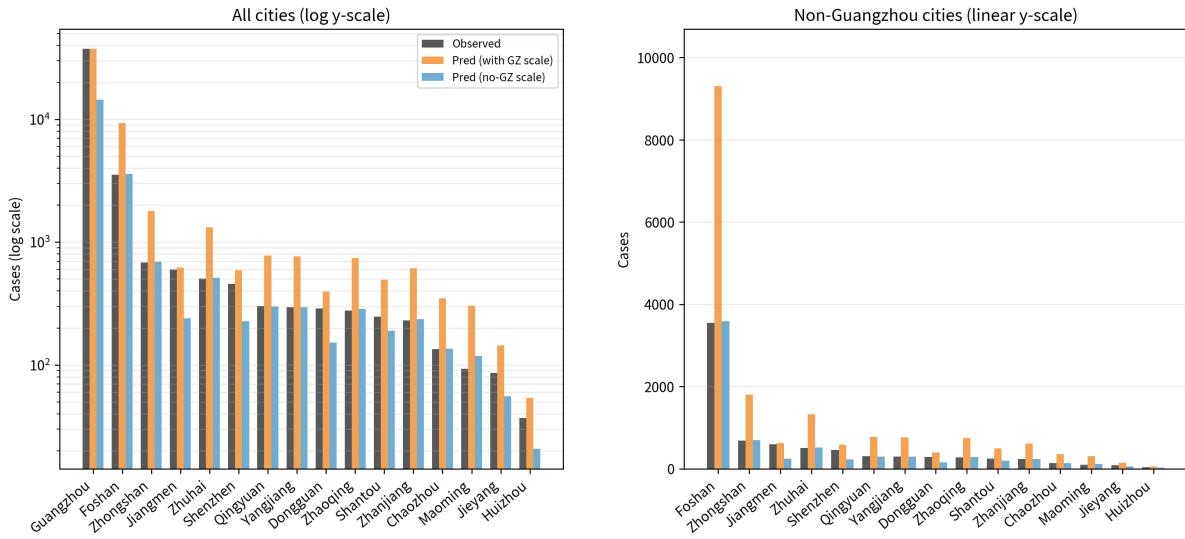


图 12: 附录 D13: 2014 年多城市年度观测与预测并列对比。

2014 multicity monthly curves: Observed vs predicted (no-GZ scaling)

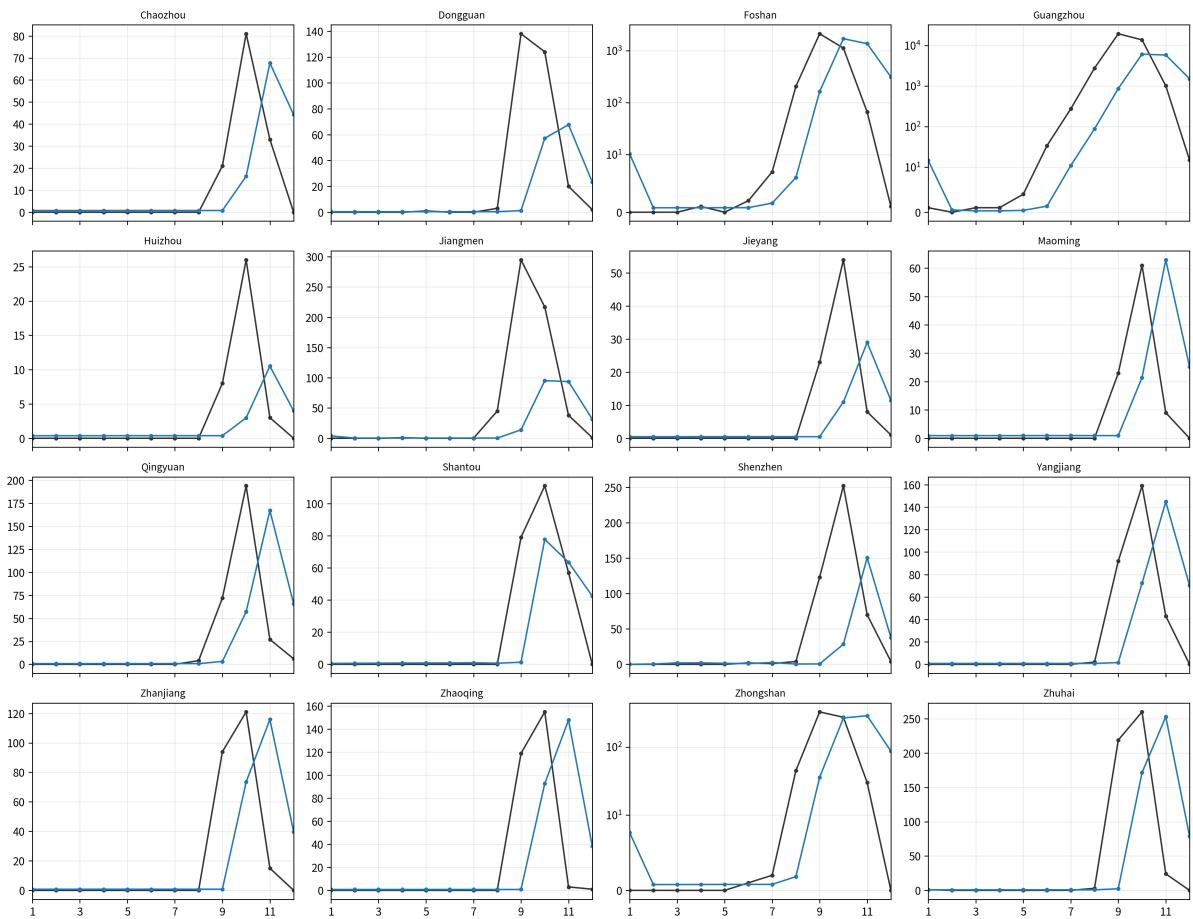


图 13: 附录 D14: 2014 年多城市月度观测-预测曲线 (去广州缩放)。

A.3 各城市逐月拟合曲线与 β' 时序（附录新增）

图 D15 给出 16 城市的总览网格图：每个城市左侧为月度观测与预测病例曲线（对数坐标），右侧为公式 $\beta'(T, H, R)$ 的时间序列。图 D16–D31 分别给出各城市的独立大图，便于逐城市审查拟合质量与传播效率的季节波动特征。

All cities: observed vs predicted cases & $\beta'(T, H, R)$ (2005–2019)



图 14: 附录 D15: 16 城市月度拟合总览——左列为观测 vs 预测病例 (对数坐标), 右列为 $\beta'(T, H, R)$ 时序 (2005–2019)。

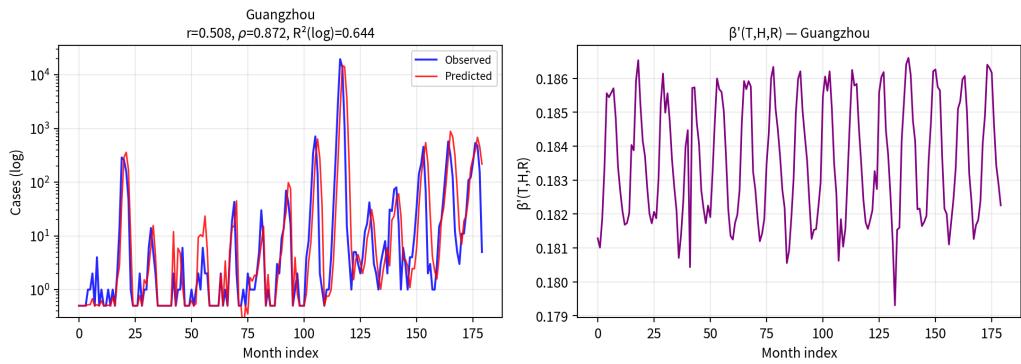


图 15: 附录 D16: 广州 (Guangzhou) 月度拟合曲线与 β' 。

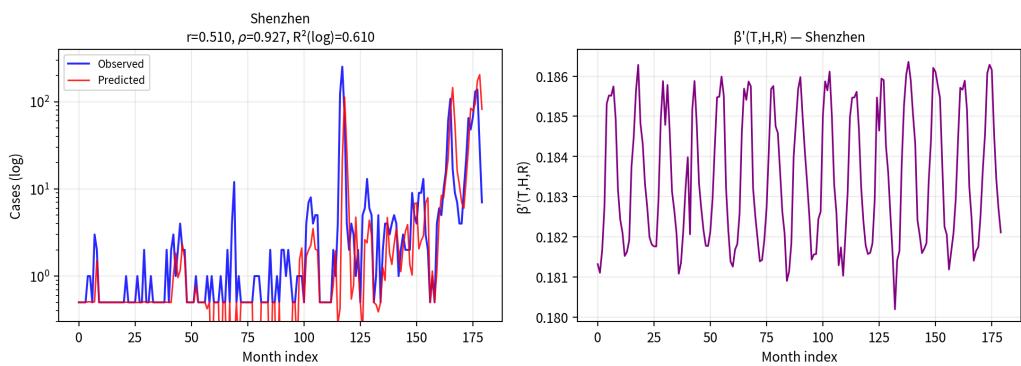


图 16: 附录 D17: 深圳 (Shenzhen) 月度拟合曲线与 β' 。

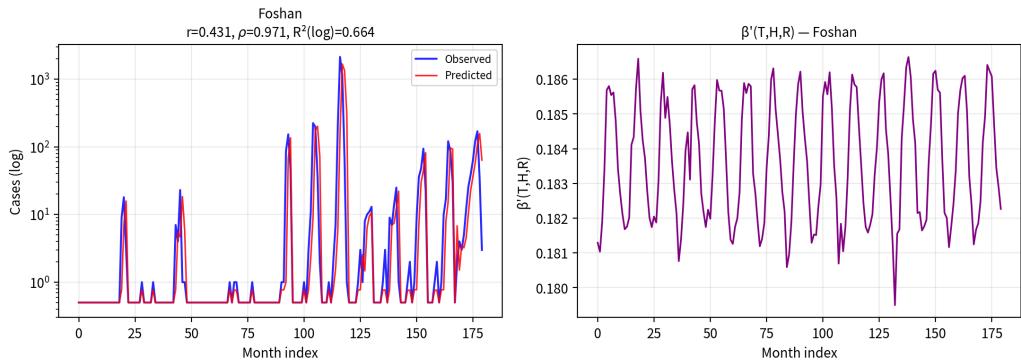


图 17: 附录 D18: 佛山 (Foshan) 月度拟合曲线与 β' 。

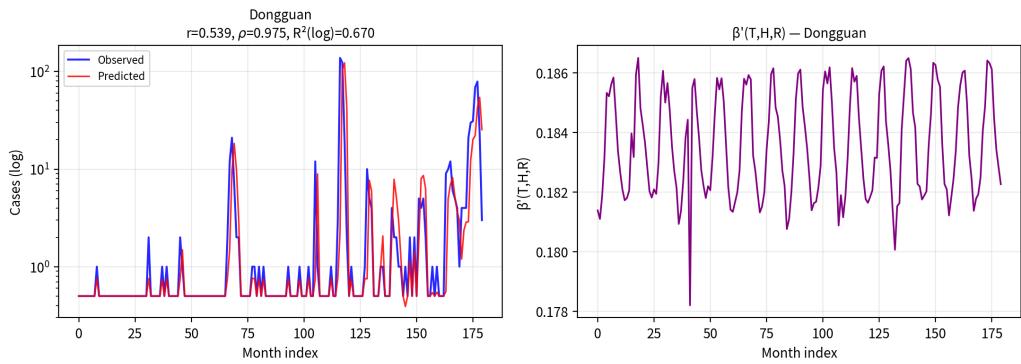


图 18: 附录 D19: 东莞 (Dongguan) 月度拟合曲线与 β' 。

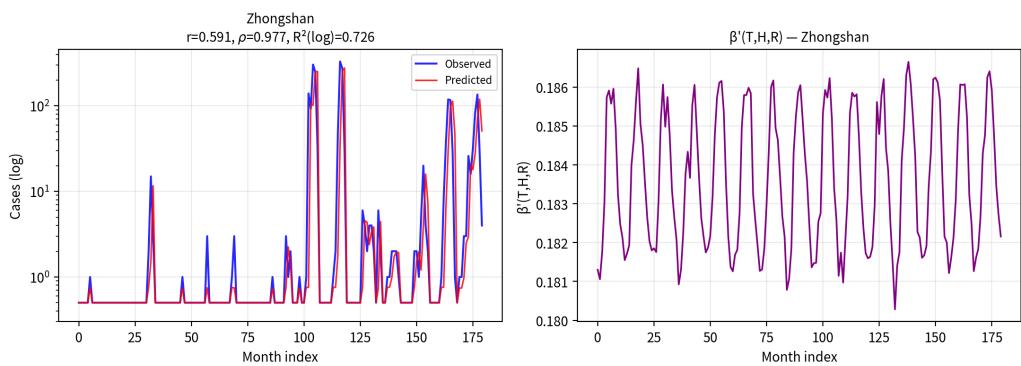


图 19: 附录 D20: 中山 (Zhongshan) 月度拟合曲线与 β' 。

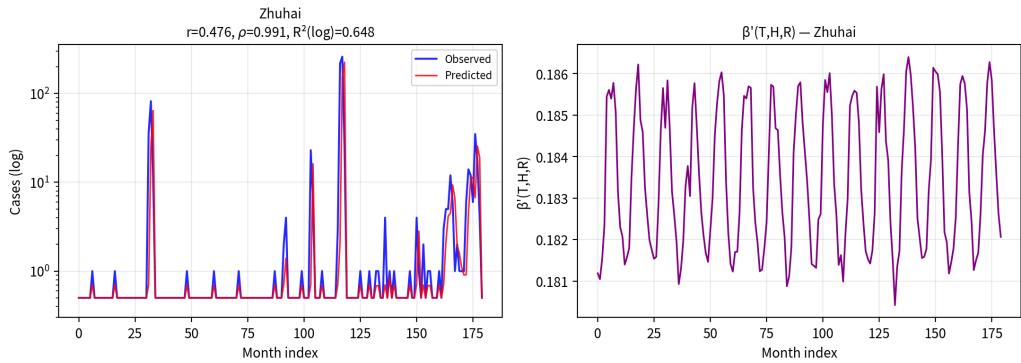


图 20: 附录 D21: 珠海 (Zhuhai) 月度拟合曲线与 β' 。

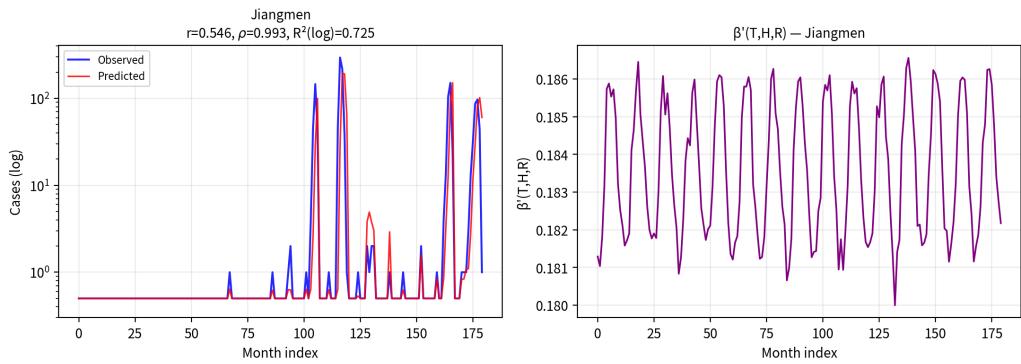


图 21: 附录 D22: 江门 (Jiangmen) 月度拟合曲线与 β' 。

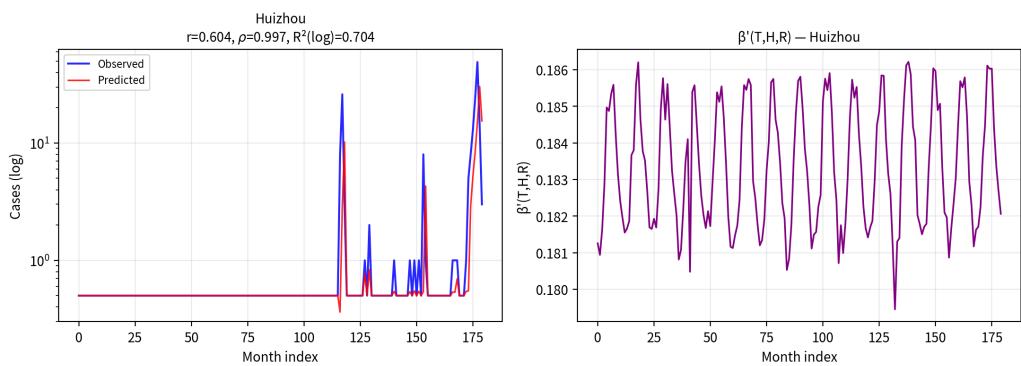


图 22: 附录 D23: 惠州 (Huizhou) 月度拟合曲线与 β' 。

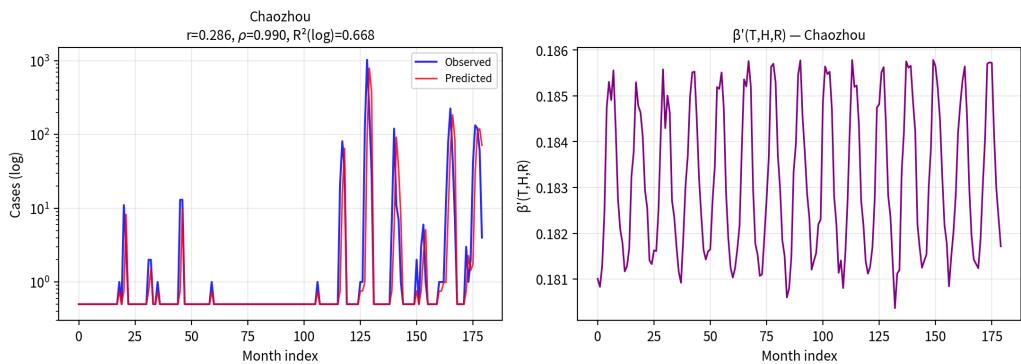


图 23: 附录 D24: 潮州 (Chaozhou) 月度拟合曲线与 β' 。

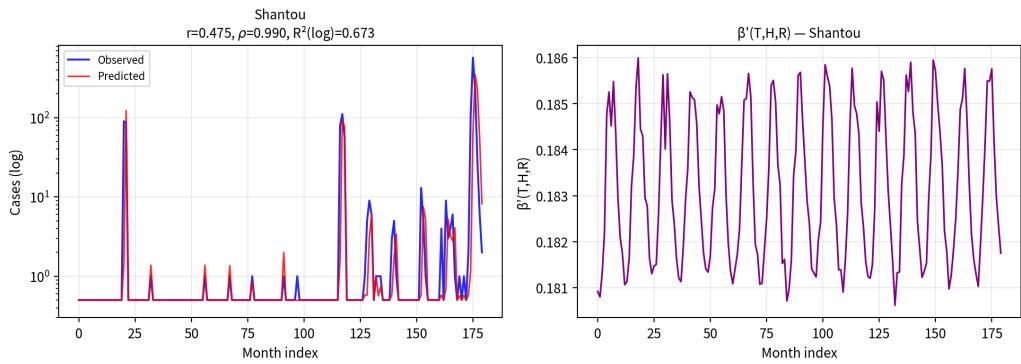


图 24: 附录 D25: 汕头 (Shantou) 月度拟合曲线与 β' 。

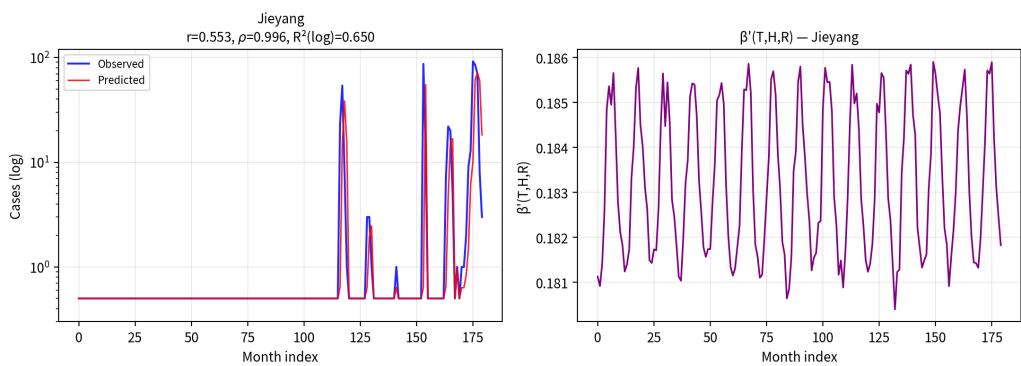


图 25: 附录 D26: 揭阳 (Jieyang) 月度拟合曲线与 β' 。

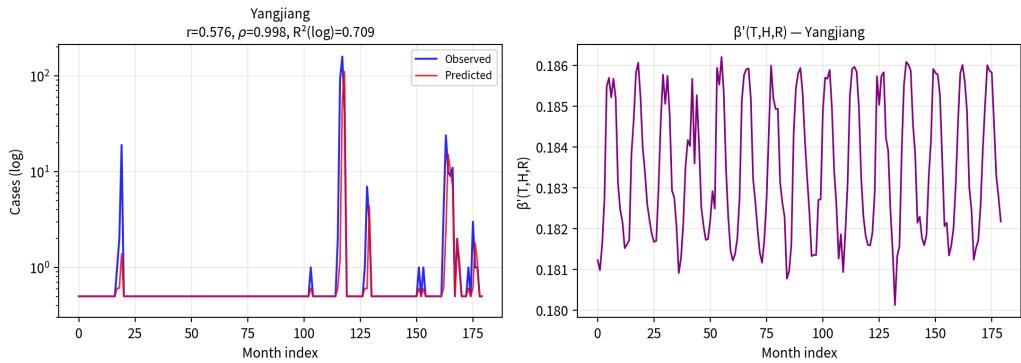


图 26: 附录 D27: 阳江 (Yangjiang) 月度拟合曲线与 β' 。

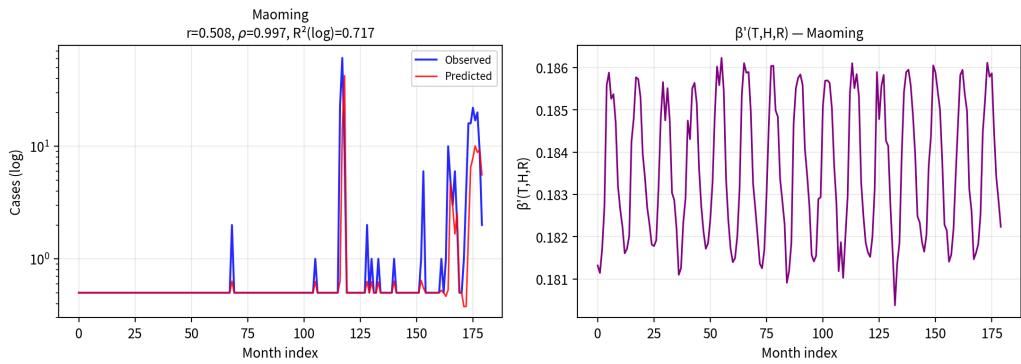


图 27: 附录 D28: 茂名 (Maoming) 月度拟合曲线与 β' 。

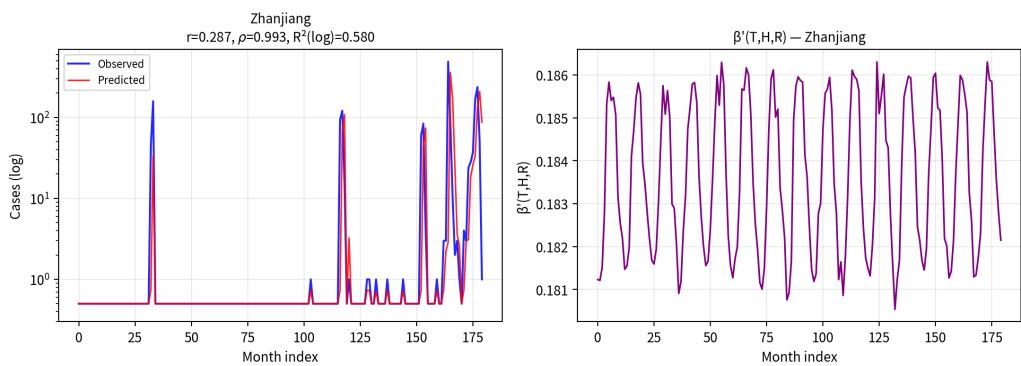


图 28: 附录 D29: 湛江 (Zhanjiang) 月度拟合曲线与 β' 。

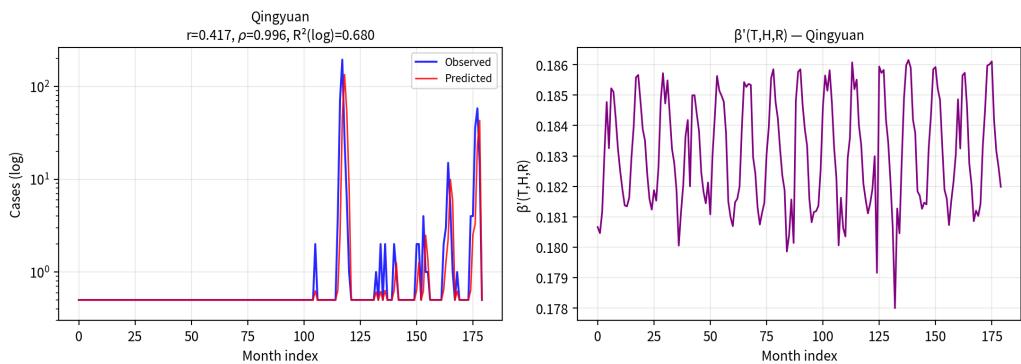


图 29: 附录 D30: 清远 (Qingyuan) 月度拟合曲线与 β' 。

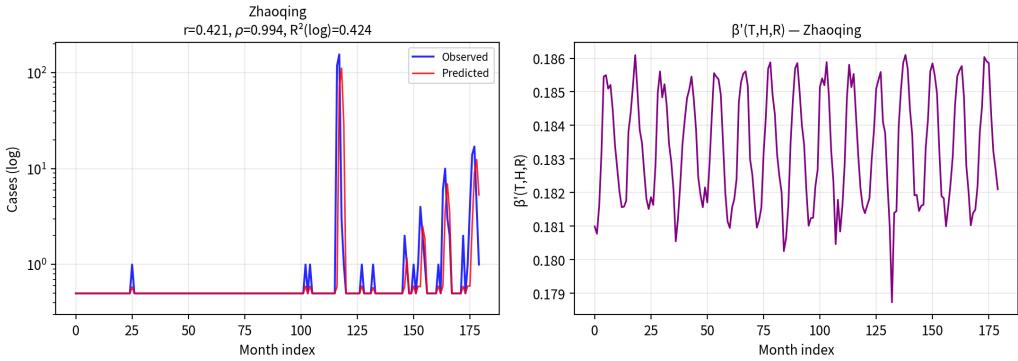


图 30: 附录 D31: 肇庆 (Zhaoqing) 月度拟合曲线与 β' 。

参考文献

- [1] World Health Organization. Dengue and severe dengue. WHO Fact Sheet, 2023.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] CCM14: Mosquito surveillance data in China. <https://github.com/xyyu001/CCM14>
- [6] 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.
- [7] 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.