

# 毕业论文

## 基于神经网络耦合动力学模型的 登革热传播效率发现与多城市验证

Discovery of Dengue Transmission Efficiency via  
Neural-Network-Coupled Dynamical Models  
and Multi-City Validation

学 院： 公共卫生学院

专 业： 流行病与卫生统计学

研究方向： 传染病建模与预测

2025 年 6 月

## 摘 要

登革热是全球最严重的蚊媒传染病之一，中国南方尤其广东省是国内最主要的流行区域。理解气候因素如何驱动蚊媒密度和登革热传播效率，对于建立早期预警系统和制定精准防控策略具有重要意义。然而，传统机制模型往往依赖先验假设固定蚊媒密度或传播率的函数形式，难以从数据中自动发现最优的气候-蚊媒关系；纯数据驱动的机器学习方法则存在可解释性不足的瓶颈。

本文提出一种“神经网络 + 符号回归 + SEIR 动力学”两部分混合建模框架，旨在兼顾机制可解释性与数据适应性。第一部分以广东省 8 个拥有布雷图指数 (BI) 监测数据的城市为研究对象 (306 条月度样本)，构建 6 维多层感知机 (MLP) 学习气候变量 (温度  $T$ 、湿度  $H$ 、降水  $R$ ) 及城市气候均值 ( $T_{\text{mean}}$ 、 $H_{\text{mean}}$ 、 $R_{\text{mean}}$ ) 到蚊媒密度  $\hat{M}$  的非线性映射 (Pearson  $r = 0.776$ ,  $R^2 = 0.597$ )，再通过知识蒸馏生成 10,000 个网格点，利用 PySR 符号回归 (Julia SymbolicRegression.jl) 从 Pareto 前沿中发现可解释的闭合公式。以真实数据  $R^2$  为选择准则，最优公式 (复杂度 25) 在真实数据上达到  $R^2 = 0.438$ 、 $r = 0.692$ 。公式结构表明蚊媒密度主要由月温度  $T$ 、城市平均温度  $T_{\text{mean}}$  和城市平均降水  $R_{\text{mean}}$  驱动，城市气候均值作为城市固定效应的代理变量有效捕捉了跨城市 BI 基线差异。留一城市交叉验证 (8 折) 的平均  $r = 0.586 \pm 0.217$ ，证实公式的跨城市泛化能力。

第二部分将公式代入 SEIR 动力学模型，结合 Brière 温度依赖传播率  $\beta'(T) = c \cdot T \cdot (T - T_{\text{min}}) \cdot \sqrt{T_{\text{max}} - T}$ ，仅凭气象数据预测登革热月度病例。在广州市数据上通过差分进化算法优化 4 个参数，得到最优温度区间  $T_{\text{min}} = 18.0^\circ\text{C}$ 、 $T_{\text{max}} = 33.0^\circ\text{C}$  (最优传播温度  $T_{\text{opt}} \approx 27^\circ\text{C}$ )，广州训练集  $R_{\text{log}}^2 = 0.448$ 、Spearman  $\rho = 0.648$ 。将模型迁移至广东省 16 个地级市，拥有 BI 数据的 8 城平均 Spearman  $\rho = 0.411$ ，无 BI 数据的 8 城  $\rho = 0.328$ ，表明公式驱动的 SEIR 模型可跨城市捕捉登革热季节性趋势。

本研究的主要创新包括：(1) 提出“NN+PySR 符号蒸馏”范式，从蚊媒监测数据中自动发现气候-蚊媒密度的解析公式，克服了传统模型依赖先验函数形式的局限；(2) 引入城市气候均值作为城市固定效应代理特征，使公式具有跨城市可迁移性；(3) 发现的闭合公式揭示了月温度、城市基线温度和降水的联合驱动作用，具有明确的生态学含义；(4) 两部分框架将公式发现与动力学预测有机结合，首次在广东省 8+16 城市尺度上系统验证了公式驱动的登革热预测能力。

**关键词：**登革热；SEIR 模型；神经网络；符号回归；传播效率；广东省；多城市验证

## Abstract

Dengue fever is one of the most severe mosquito-borne infectious diseases globally. Southern China, especially Guangdong Province, is the primary endemic region in the country. Understanding how climatic factors drive mosquito density and dengue transmission efficiency is crucial for establishing early-warning systems and formulating targeted control strategies. However, traditional mechanistic models rely on *a priori* assumptions to fix the functional form of mosquito density or transmission rates, making it difficult to discover optimal climate–mosquito relationships from data automatically. Meanwhile, purely data-driven machine learning approaches suffer from limited interpretability.

This thesis proposes a two-part hybrid modeling framework—“Neural Network + Symbolic Regression + SEIR Dynamics”—that aims to balance mechanistic interpretability with data adaptability. Part I uses 8 cities in Guangdong Province with Breteau Index (BI) monitoring data (306 monthly samples) to train a 6-dimensional MLP that learns the nonlinear mapping from climate variables (temperature  $T$ , humidity  $H$ , precipitation  $R$ ) and city climate means ( $T_{\text{mean}}$ ,  $H_{\text{mean}}$ ,  $R_{\text{mean}}$ ) to normalized mosquito density  $\hat{M}$ , achieving Pearson  $r = 0.776$  and  $R^2 = 0.597$ . Knowledge distillation generates 10,000 grid points, and PySR (Julia SymbolicRegression.jl) discovers interpretable closed-form formulas from the Pareto front. Selecting by real-data  $R^2$ , the best formula (complexity 25) achieves  $R^2 = 0.438$  and  $r = 0.692$  on observed data. The formula structure reveals that mosquito density is primarily driven by monthly temperature  $T$ , city mean temperature  $T_{\text{mean}}$ , and city mean precipitation  $R_{\text{mean}}$ , where city climate means serve as proxy variables for city-level fixed effects that effectively capture cross-city BI baseline differences. Leave-one-city-out cross-validation (8 folds) yields mean  $r = 0.586 \pm 0.217$ , confirming the formula’s cross-city generalizability.

Part II integrates the discovered formula into an SEIR compartmental model with a Brière temperature-dependent transmission rate  $\beta'(T) = c \cdot T \cdot (T - T_{\min}) \cdot \sqrt{T_{\max} - T}$ , enabling dengue case prediction from climate data alone. Differential evolution optimizes four parameters on Guangzhou data, yielding an optimal transmission temperature range of  $T_{\min} = 18.0^\circ\text{C}$  to  $T_{\max} = 33.0^\circ\text{C}$  ( $T_{\text{opt}} \approx 27^\circ\text{C}$ ). The model achieves training  $R_{\log}^2 = 0.448$  and Spearman  $\rho = 0.648$  for Guangzhou. Transferring to 16 prefecture-level cities, the 8 cities with BI data achieve mean Spearman  $\rho = 0.411$ , while the 8 cities without BI data achieve  $\rho = 0.328$ , demonstrating the formula-driven SEIR model’s ability to capture dengue seasonal trends across cities.

Key innovations include: (1) an “NN + PySR symbolic distillation” paradigm that automatically discovers climate–mosquito density analytical formulas from monitoring data, overcoming the reliance on *a priori* functional forms; (2) the introduction of city climate means as proxy features for city-level fixed effects, enabling cross-city transferability; (3) a

discovered closed-form formula revealing the joint driving role of monthly temperature, city baseline temperature, and precipitation, with clear ecological interpretation; (4) a two-part framework that organically combines formula discovery with dynamical prediction, systematically validated at the 8+16 city scale in Guangdong Province.

**Keywords:** Dengue fever; SEIR model; Neural network; Symbolic regression; Transmission efficiency; Guangdong Province; Multi-city validation

目录

|                                       |    |
|---------------------------------------|----|
| 摘要                                    | I  |
| Abstract                              | II |
| 前言                                    | 1  |
| 1 第I部分 气候驱动蚊媒密度公式发现——基于神经网络与符号回归的知识蒸馏 | 5  |
| 1.1 引言                                | 5  |
| 1.2 数据材料和方法                           | 5  |
| 1.2.1 研究数据来源与数据预处理                    | 5  |
| 1.2.2 神经网络架构                          | 6  |
| 1.2.3 训练策略与损失函数                       | 6  |
| 1.2.4 知识蒸馏与符号回归                       | 7  |
| 1.2.5 留一城市交叉验证                        | 7  |
| 1.2.6 模型评价指标                          | 7  |
| 1.3 结果                                | 7  |
| 1.3.1 8城蚊媒密度与气象因素基本特征                 | 7  |
| 1.3.2 神经网络拟合结果                        | 8  |
| 1.3.3 PySR 符号回归公式发现                   | 9  |
| 1.3.4 留一城市交叉验证                        | 10 |
| 1.4 讨论                                | 11 |
| 1.5 本章小结                              | 12 |
| 2 第II部分 公式驱动的 SEIR 登革热预测——多城市迁移与验证    | 12 |
| 2.1 引言                                | 12 |
| 2.2 数据材料和方法                           | 13 |
| 2.2.1 多城市数据概况                         | 13 |
| 2.2.2 SEIR 动力学模型                      | 13 |
| 2.2.3 参数优化                            | 13 |
| 2.2.4 跨城市迁移与分组比较                      | 14 |
| 2.2.5 评估指标                            | 14 |
| 2.3 结果                                | 14 |
| 2.3.1 广州 SEIR 参数优化                    | 14 |
| 2.3.2 16城迁移与分组比较                      | 15 |
| 2.3.3 公式 $\hat{M}$ 与实测 BI 比较          | 17 |

|                      |           |
|----------------------|-----------|
| 目录                   | V         |
| 2.4 讨论               | 18        |
| 2.5 本章小结             | 19        |
| <b>3 总结与展望</b>       | <b>19</b> |
| 研究总结                 | 19        |
| 研究创新点                | 20        |
| 研究展望                 | 20        |
| <b>参考文献</b>          | <b>21</b> |
| <b>附录一、论文涉及的图表补充</b> | <b>28</b> |
| <b>致谢</b>            | <b>29</b> |

## 前 言

登革热 (Dengue Fever) 是由登革病毒 (DENV) 引起、主要经由伊蚊 (*Aedes aegypti* 和 *Aedes albopictus*) 叮咬传播的急性虫媒传染病。近年来,随着全球气候变暖、城市化进程加速以及国际贸易和旅游的频繁,登革热已成为世界上增长最快的虫媒病毒性疾病<sup>[1]</sup>。据 Bhatt 等<sup>[2]</sup> 估计,全球每年约有 3.9 亿人感染登革病毒,其中约 9600 万例出现临床症状。世界卫生组织<sup>[3]</sup> 指出,过去二十年间登革热报告病例数增长了八倍以上,从 2000 年的 50 万例升至 2023 年的超过 600 万例,现已在 100 多个国家流行。Messina 等<sup>[1]</sup> 利用全球尺度的统计模型预测,到 2080 年气候变化和城市化将使全球约 63 亿人面临登革热风险,较 2015 年增加约 22 亿人。

在中国,登革热虽不是本土地方性流行病,但自 20 世纪 70 年代末以来,由输入性病例引发的本土暴发在东南沿海地区频发。特别是广东省,地处亚热带,气候温暖湿润,极适宜白纹伊蚊的生长繁殖,长期以来是我国登革热防控的重点区域<sup>[5, 4]</sup>。2014 年,广东省经历了历史上最严重的登革热疫情,报告病例数超过 45,000 例,广州市单城报告逾 37,000 例,创下历史纪录<sup>[6]</sup>。Yue 等<sup>[4]</sup> 的系统综述表明,自 2004 年以来广东省贡献了全国超过 70% 的登革热报告病例,年度病例数呈波动性上升趋势,暴发间隔呈缩短趋势。这一流行模式与该地区亚热带季风气候、高度城市化、人口密集以及频繁的国际人员流动密切相关。从血清型分布来看,广东省历年暴发中 DENV-1 最为常见,但也检测到 DENV-2、DENV-3 和 DENV-4 的输入性和本地传播病例<sup>[6]</sup>。值得注意的是,由于不同血清型之间仅存在短暂的交叉免疫保护,二次感染可能导致更严重的登革出血热 (DHF) 和登革休克综合征 (DSS),给公共卫生系统带来额外压力<sup>[7]</sup>。近年来,全球气候变暖和极端天气事件增加,进一步加剧了登革热北扩和暴发频次增加的风险。DeSouza 等<sup>[8]</sup> 指出,2023–2024 年全球登革热病例再创历史新高,部分与厄尔尼诺现象引发的异常高温和强降水有关。因此,深入探究登革热的传播机制,特别是量化环境因素对传播过程的非线性驱动作用,对于制定精准的防控策略具有重要的现实意义。

## 气候因素与蚊媒传播

气候因素是驱动蚊媒传染病时空分布和流行强度的核心外部变量。伊蚊的生命周期、种群密度及病毒在蚊体内的复制速率均受到气象条件的严格制约<sup>[7]</sup>。**温度**直接影响蚊虫的生殖周期、幼虫发育率及成蚊存活率<sup>[8, 9, 10]</sup>。更重要的是,温度决定了外潜伏期 (Extrinsic Incubation Period, EIP),即病毒在蚊体内复制并具备传播能力所需的时间<sup>[10, 11]</sup>。Mordecai 等<sup>[12]</sup> 的全面实验研究表明,蚊媒传播能力对温度呈单峰响应,最优传播温度约为 29°C。在此温度下,蚊虫叮咬率最高、病毒外潜伏期最短、蚊虫存活率最大,三者的乘积效应使传播效率达到峰值。当温度低于约 18°C 或高于约 34°C 时,传播能力显著下降<sup>[34]</sup>。Shapiro 等<sup>[9]</sup> 和 Lambrechts 等<sup>[10]</sup> 进一步指出,温度日较差 (Diurnal Temperature Range, DTR) 对传播效率也有重要影响。Colón-González 等<sup>[13]</sup> 基

于多模型集合预测发现,温度的升高将显著扩大登革热的适传播区,若全球升温幅度能控制在  $2.0^{\circ}\text{C}$ ,可避免拉丁美洲每年约 280 万例新增登革热病例。Kamiya 等<sup>[11]</sup>的荟萃分析确认了温度对蚊媒传染病传播的非线性调控作用在全球不同地理区域具有一致性。

**降水**对登革热的影响具有双重性。一方面,降水为蚊虫提供了必要的繁殖栖息地——积水容器、洼地和废弃物中的积水是伊蚊的主要产卵场所<sup>[16]</sup>;适量降水显著增加蚊虫密度,从而提高传播风险<sup>[15]</sup>。另一方面,极端强降水可能冲刷幼虫栖息地、降低蚊虫存活率,产生抑制效应<sup>[13]</sup>。Zhou 等<sup>[14]</sup>的纵向研究发现,降水与登革热发病率之间存在显著的非线性关系和时间滞后效应,累积降水量超过一定阈值后传播风险不再持续增加,呈现饱和或下降趋势。Cheng 等<sup>[17]</sup>针对广州的研究发现,在前期水分充足的条件下,滞后 7–121 天的强降雨会降低登革热风险。这种“先增后平”的模式在本文模型发现中也得到了印证。

**相对湿度**影响蚊虫的存活和活动能力。Wu 等<sup>[19]</sup>对中国南方登革热暴发的时间序列分析发现,相对湿度存在一个约 76% 的阈值效应——当湿度超过此值时,蚊虫存活率和叮咬活跃度显著提高,登革热传播风险明显增大。Cheng 等<sup>[17]</sup>对广州的研究进一步证实了湿度与登革热发病率之间的正相关关系,尤其在高温环境下湿度的促进作用更为显著。Polrob 等<sup>[18]</sup>在东南亚的研究中发现,湿度与蚊虫叮咬率之间存在协同关系,进一步支持了湿度作为重要传播调节因子的地位。

从生态机制的角度,上述三个气候变量并非独立作用,而是通过复杂的交互效应共同决定传播强度。例如,高温高湿条件下蚊虫的吸血频率和存活率同时增加,产生协同促进效应;而高温干燥条件则可能因蚊虫脱水死亡而抑制传播。DaCosta 等<sup>[20]</sup>和 Leung 等<sup>[21]</sup>的研究均强调,单独考虑任一气候因子都不足以准确描述传播动态,需要同时纳入温度、降水和湿度的联合效应。这一认识构成了本文将三个气候变量同时纳入神经网络模型的理论基础。

## 登革热建模研究现状

登革热建模研究经历了从纯统计模型到机制模型、再到人工智能融合模型的发展历程,不同方法在解释能力、预测精度和可推广性方面各有优劣。

**统计模型。**广义加性模型 (GAM) 和分布式滞后非线性模型 (DLNM) 是登革热气候–疫情关系研究中应用最广泛的统计工具<sup>[62]</sup>。GAM 能够灵活地刻画气候变量与发病率之间的非线性关系,同时控制季节性和长期趋势等混杂因素。DLNM 进一步考虑了气候影响的时间滞后结构,能够同时估计暴露–反应关系和滞后效应<sup>[63]</sup>。Liu 等<sup>[22]</sup>利用 DLNM 分析了中国南方多个城市的气候–登革热关系,发现温度和降水的影响在滞后 1–3 个月最为显著。Luo 等<sup>[24]</sup>对马来西亚、新加坡和泰国 2017–2022 年的登革热传播模式进行研究,发现最高气温与登革热关系的峰值相对风险在 COVID-19 后显著上升。Cheng 等<sup>[25]</sup>基于中国广东和浙江 2005–2024 年的数据,构建了融合 DLNM



与混合智能算法的预测框架，在平均准确率等指标上均表现最优。然而，统计模型本质上是“关联性”而非“因果性”工具，其参数不具有直接的流行病学机制含义，在外推到未见过的的气候条件或新的地理区域时，预测能力往往大幅下降<sup>[26, 27]</sup>。

**机制模型。**基于仓室结构的传染病动力学模型是理解传播机制的经典工具。Ross-Macdonald 模型及其扩展形式将人-蚊传播过程分解为若干关键参数，每个参数都具有明确的生物学含义<sup>[29]</sup>。SEI-SEIR 耦合模型是登革热研究中常用的仓室结构，将蚊群的“易感-暴露-感染”与人群的“易感-暴露-感染-恢复”动态耦合<sup>[30]</sup>。Li 等<sup>[37]</sup>在 2019 年 *PNAS* 上发表的研究中，在 SEI-SEIR 框架中使用时变三次样条函数拟合传播系数  $\beta(t)$  与温度的关系，并通过广州 2005–2015 年的病例和气候数据进行参数估计，发现  $\beta(t)$  对温度呈单峰响应，最优温度约为 27–29°C。然而，该方法存在以下局限：(1) 三次样条的形式需要预先指定节点数和位置；(2) 仅考虑温度单一气候变量，忽略了降水和湿度；(3) 最终结果为分段平滑曲线而非可移植的闭合公式。Caldwell 等<sup>[36]</sup>指出，实验室环境与复杂的野外环境存在巨大差异，直接套用实验室参数往往导致模型预测偏差。现有的机制模型通常直接采用实验室测定的温依参数（如 Brière 函数描述叮咬率）<sup>[34, 51]</sup>，难以真实反映野外条件下气候因素对传播效率的综合影响。

**人工智能融合方法。**近年来，将深度学习与微分方程模型结合的“物理信息神经网络” (PINN) 和“神经常微分方程” (Neural ODE) 方法受到越来越多的关注<sup>[61]</sup>。在传染病建模领域，Sehi 等<sup>[23]</sup>和 Luo 等<sup>[24]</sup>将 PINN 应用于 SIR/SEIR 模型的参数估计和短期预测，取得了优于传统拟合方法的精度。Li 等<sup>[40]</sup>将 COVID-19 模型动态嵌入物理信息神经网络，同时推断未知参数和未观察到的底层模型动态。Nikparvar 等<sup>[41]</sup>将人口流动性作为变量输入 LSTM 用于预测美国各县的确诊病例数和死亡人数。Murphy 等<sup>[42]</sup>利用不同传染动力学生成的数据训练了一个图神经网络。然而，纯神经网络方法的“黑箱”本质使其难以提供机制层面的洞见<sup>[43]</sup>。即使模型预测准确，研究者仍然无法回答“气候如何影响传播率”这一核心科学问题。Baker 等<sup>[26]</sup>和 Mills 等<sup>[27]</sup>均指出，在传染病动力学领域，可解释性和可迁移性通常比单纯的预测精度更有实际价值。Ahman 等<sup>[28]</sup>的综述进一步强调了“混合机制-数据驱动”框架在传染病建模中的前景。Kamyshnyi 等<sup>[44]</sup>和 Adeoye 等<sup>[45]</sup>的综述也表明，神经网络虽然能捕捉复杂的非线性模式，却无法揭示疾病传播的内在动力学规律，更无法转化为可推广的数学知识。

**符号回归方法。**符号回归 (Symbolic Regression, SR) 是一种从数据中直接搜索数学表达式的方法，能够在不预设函数形式的前提下发现数据中的数学规律<sup>[59]</sup>。与传统回归方法不同，符号回归的搜索空间包含所有可能的数学表达式，其目标是在精度和复杂度之间取得帕累托最优。Fajardo 等<sup>[47]</sup>在 *PLOS Computational Biology* 发表的工作提出了贝叶斯符号回归方法，用于从报告病例和检测率数据中自动学习传染病发病率的闭式数学模型。Zhang 等<sup>[48]</sup>在 2024 年 *PLOS Computational Biology* 发表了将符号回归应用于传染病模型参数发现的开创性工作——通过将蚊媒种群动力学模型

耦合神经网络，有效揭示了伊蚊产卵率和温度、降水之间的关系，并使用符号回归确定最优函数表达式。然而，该方法面临以下挑战：(1) 直接在高维表达式空间中搜索计算成本极高；(2) 缺乏利用先验物理知识引导搜索的机制；(3) 尚未在真实登革热传播效率发现上得到充分验证。Makke 和 Mahesh<sup>[46]</sup> 在符号回归综述中指出，结合神经网络预训练和符号蒸馏的两阶段策略是一种有前景的方向：先用神经网络捕获复杂映射关系，再用符号回归提取简洁公式。这种“知识蒸馏”思路正是本文方法论的核心灵感来源。然而，目前尚未有研究将“神经网络嵌入 + 符号回归”的完整框架应用于登革热传播效率反演与公式推导中。

## 研究目标与创新点

基于上述文献回顾，本文提出以下研究目标：(1) 构建“SEIR 动力学反演 + 神经网络 + 符号蒸馏”三阶段混合建模框架，从时间序列数据中自动发现气候变量到登革热传播系数  $\beta'$  的最优函数关系。(2) 以广州市为核心案例，利用 2005–2019 年月度病例和气候数据，通过 SEIR 逆问题求解反演  $\beta'(t)$ ，训练耦合模型并提取可解释闭合公式。(3) 将发现的公式迁移至广东省 16 个地级市，在空间维度上验证其泛化能力和可迁移性。(4) 与现有方法（尤其是 Li 等 2019 年 PNAS 的样条方法和 Zhang 等 2024 年的纯符号回归方法）进行比较，论证本框架在可解释性–泛化性平衡方面的优势。

与现有工作相比，本文的创新点包括：(1) **方法论创新——“NN 逆问题 + 符号蒸馏”范式**：不同于 Li 等<sup>[37]</sup> 预设样条函数形式，本文通过神经网络自由学习  $\beta'$  的气候映射关系，再用符号回归提取公式，实现了“数据驱动的函数发现”。(2) **多变量联合建模**：不同于仅考虑温度单一变量的传统做法，本文同时纳入温度、降水和相对湿度三个气候变量及其交互效应。(3) **可迁移的闭合公式**：符号回归发现的二次多项式公式具有明确的系数含义，可直接通过城市尺度参数进行迁移，无需在每个城市重新训练模型。(4) **系统性的空间验证**：通过 16 城年度排名和月度曲线的双重验证，首次在中国南方多城市尺度上系统评估了单城市发现的传播效率公式的空间泛化性能。

## 全文结构

本文其余部分组织如下：第一部分（气候 → 蚊媒密度公式发现）以广东省 8 个有 BI 监测数据的城市为研究对象，详细阐述数据来源、神经网络架构、知识蒸馏策略、PySR 符号回归方法以及评估指标体系，呈现 NN 拟合、公式发现和留一城市交叉验证的结果与讨论。第二部分（公式驱动的 SEIR 登革热预测）将公式代入 Brière+SEIR 框架，在广州优化参数后迁移至 16 个地级市，呈现分组比较和月度验证结果。最后一章（总结与展望）总结主要发现和创新点，讨论研究局限性，提出未来改进方向。

# 1 第 I 部分 气候驱动蚊媒密度公式发现——基于神经网络与符号回归的知识蒸馏

## 1.1 引言

蚊媒密度是登革热传播动态中的关键变量——布雷图指数 (Breteau Index, BI) 作为标准化的蚊媒密度监测指标, 直接反映伊蚊孳生水平<sup>[66]</sup>。然而, BI 的实测数据在时间和空间上均覆盖有限, 制约了基于蚊媒密度的登革热预测模型向更广范围推广。如果能够从气象数据中发现蚊媒密度的气候驱动公式, 则可在缺乏实测 BI 的城市中仅凭气象数据估算蚊媒密度, 从而为 SEIR 等动力学模型提供关键输入。

现有研究主要依赖实验室参数 (如 Brière 方程) 描述温度对蚊虫种群的影响<sup>[51, 34]</sup>, 但实验室恒温环境难以真实反映野外复杂的多因素联合作用<sup>[53, 36]</sup>。Li 等<sup>[37]</sup> 使用 GAM 样条拟合气象-蚊媒关系, 但其结果为分段平滑曲线, 不具有闭合公式形式, 难以迁移至其他城市。Zhang 等<sup>[48]</sup> 将符号回归应用于蚊媒种群动力学, 但直接在高维表达式空间中搜索计算成本极高。

本章利用广东省 8 个拥有 BI 监测数据的城市 (306 条月度样本), 构建“神经网络 + 符号回归”框架, 旨在回答一个关键问题: 气温、湿度和降雨通过什么样的数学关系驱动蚊媒密度? 本章首先训练多层感知机 (MLP) 学习气象变量到蚊媒密度的非线性映射, 再通过 PySR 符号回归从神经网络中蒸馏出可解释的闭合公式。为捕捉不同城市间 BI 基线的巨大差异 (如深圳均值 2.2 vs 揭阳均值 13.1), 本章引入城市气候均值作为代理特征, 使公式具有跨城市泛化能力。

选择广东省 8 城作为研究对象的理由包括: (1) 这 8 个城市拥有较完整的 BI 监测记录, 覆盖 2016–2019 年; (2) 城市间气候条件和 BI 水平存在有意义的梯度差异, 为公式发现提供了充分的信号变异性; (3) 留一城市交叉验证可严格检验公式的空间泛化能力。

## 1.2 数据材料和方法

### 1.2.1 研究数据来源与数据预处理

本章选取广东省 8 个拥有布雷图指数 (BI) 监测数据的城市作为研究对象: 东莞、广州、惠州、江门、揭阳、茂名、汕头和深圳。这 8 个城市覆盖了珠三角核心区 (广州、深圳、东莞)、珠三角外围区 (惠州、江门) 及粤东 (揭阳、汕头) 和粤西 (茂名), 在地理位置和气候条件上具有较好的代表性。

**气象数据:** 包括月平均气温 ( $T$ , °C)、月平均相对湿度 ( $H$ , %) 及月累计降雨量 ( $R$ , mm), 来源于美国国家海洋和大气管理局 (NOAA) 下属的国家环境信息中心 (NCEI)。

**蚊媒监测数据:** 来源于广东省疾病预防控制中心, 使用布雷图指数 (Breteau Index, BI), 即每 100 户居民中发现孳生伊蚊幼虫的积水容器数<sup>[66]</sup>。合并后共获得 306 条月

度样本，各城市样本量为：东莞 24 条、广州 87 条、惠州 9 条、江门 36 条、揭阳 18 条、茂名 18 条、汕头 44 条、深圳 70 条。

**数据预处理。**为消除量纲差异，对 6 维输入特征进行 Min-Max 标准化：

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

目标变量采用  $\log(1 + \text{BI})$  变换，保留跨城市的绝对信号差异。

**城市气候均值特征。**为捕捉不同城市间 BI 基线的巨大差异（深圳均值 2.2 vs 揭阳均值 13.1），引入城市气候均值作为代理特征：

$$T_{\text{mean},c} = \frac{1}{n_c} \sum_{t=1}^{n_c} T_{c,t}, \quad H_{\text{mean},c} = \frac{1}{n_c} \sum_{t=1}^{n_c} H_{c,t}, \quad R_{\text{mean},c} = \frac{1}{n_c} \sum_{t=1}^{n_c} R_{c,t} \quad (2)$$

其中  $c$  为城市索引， $n_c$  为该城市的样本数。城市气候均值是该城市气候基线的代理变量，其物理含义在于：公式同时捕捉”季节性波动”（如  $T - T_{\text{mean}}$ ）和”城市基线”（ $T_{\text{mean}}$  本身），从而在无需显式城市标签的条件下实现跨城市泛化。

### 1.2.2 神经网络架构

为学习气候变量到蚊媒密度  $\hat{M}$  的非线性映射，构建 6 维输入的 MLP：

$$\hat{M}(T, H, R, T_{\text{mean}}, H_{\text{mean}}, R_{\text{mean}}; \theta) = \text{MLP}(\mathbf{x}; \theta) \quad (3)$$

网络架构为：输入层 6 维，隐藏层 1 为 64 个神经元（Softplus 激活， $f(x) = \ln(1 + e^x)$ ，Dropout 0.1），隐藏层 2 为 64 个神经元（Softplus，Dropout 0.1），隐藏层 3 为 32 个神经元（Softplus），输出层 1 个神经元（Softplus，保证  $\hat{M} \geq 0$ ）。总参数量为  $6 \times 64 + 64 + 64 \times 64 + 64 + 64 \times 32 + 32 + 32 \times 1 + 1 = 6,753$ 。

Softplus 激活函数保证了输出的光滑性和非负性。Dropout 0.1 用于防止过拟合（306 样本量相对较小）。输出层直接预测  $\log(1 + \text{BI})$ ，保留了跨城市的绝对信号差异。

### 1.2.3 训练策略与损失函数

损失函数设计为 Huber 损失<sup>[69]</sup>与 Pearson 相关系数的加权组合：

$$\mathcal{L}(\theta) = \text{Huber}(\hat{M}, \log(1 + \text{BI})) + 0.3 \cdot (1 - r(\hat{M}, \log(1 + \text{BI}))) \quad (4)$$

Huber 损失（ $\delta = 1.0$ ）对 BI 分布右偏导致的极端值更鲁棒；Pearson 相关项鼓励网络捕捉季节性变化趋势。

参数更新采用 Adam 优化器<sup>[58]</sup>，初始学习率  $3 \times 10^{-3}$ ，权重衰减  $10^{-4}$ ，结合余弦退火（Cosine Annealing）学习率调度策略，共训练 5,000 轮。采用早停策略，保存验证损失最低的模型权重。



#### 1.2.4 知识蒸馏与符号回归

以训练好的神经网络为“教师”，采用知识蒸馏策略生成高密度虚拟数据集：在 6 维输入空间中，对  $T/H/R$  各取 20 个等距网格点，对  $T_{\text{mean}}/H_{\text{mean}}/R_{\text{mean}}$  各取 5 个分位数点，通过拉丁超立方采样生成 10,000 个网格点，以神经网络预测值为目标。

利用 Julia 语言的 SymbolicRegression.jl (PySR 后端) 在蒸馏数据上搜索最优公式<sup>[59]</sup>。搜索配置为：二元运算符  $\{+, -, \times, /\}$ ，一元运算符  $\{\exp, \cos, \sqrt{\cdot}\}$ ，最大复杂度 25，种群数 30，迭代 300 轮，超时 300 秒。变量名为  $T, H, R, T_m, H_m, R_m$ 。

PySR 返回 Pareto 前沿上的一系列候选公式（复杂度-精度权衡）。与传统做法选取蒸馏 loss 最低的公式不同，本文提出以**真实数据**  $R^2$  为选择准则：对每个候选公式，在 306 条真实数据上计算  $R^2$ ，选择  $R^2$  最高者。这一策略有效避免了过拟合蒸馏网格的问题。

#### 1.2.5 留一城市交叉验证

采用留一城市交叉验证 (Leave-One-City-Out CV, LOCO CV) 评估 NN 的跨城市泛化能力：每折留出一个城市作为测试集，用其余 7 个城市训练 NN (2,000 轮，学习率  $5 \times 10^{-3}$ )，在测试城市上评估预测性能。8 折 LOCO CV 的均值和标准差反映公式的空间泛化能力。

#### 1.2.6 模型评价指标

采用多维度评价指标体系：(1) **线性相关**：Pearson 相关系数  $r$  (首要指标)。(2) **拟合优度**：决定系数  $R^2 = 1 - \sum(\hat{M}_i - M_i)^2 / \sum(M_i - \bar{M})^2$ 。(3) **误差指标**：均方根误差 RMSE。(4) **信息准则**：AIC 和 BIC 用于公式复杂度-精度权衡<sup>[73]</sup>。公式选择以真实数据  $R^2$  为首要指标，在  $R^2$  接近时优先选择更简单的公式（奥卡姆剃刀原则）。

### 1.3 结果

#### 1.3.1 8 城蚊媒密度与气象因素基本特征

合并后的数据集包含 306 条月度样本，覆盖 8 个城市。各城市的 BI 水平和气候基线差异显著（表 1）：城市平均温度从东莞、深圳的  $23.4^\circ\text{C}$  到揭阳的  $26.9^\circ\text{C}$ ，BI 均值从深圳的 2.2 到揭阳的 13.1，跨度近 6 倍。这种城市间的巨大差异构成了跨城市泛化的核心挑战，也是引入城市气候均值特征的直接动因。

表 1: 8 城数据概况

| 城市 | 样本数 | $T_{\text{mean}} (^{\circ}\text{C})$ | BI 均值 |
|----|-----|--------------------------------------|-------|
| 东莞 | 24  | 23.4                                 | 2.3   |
| 广州 | 87  | 24.6                                 | 6.2   |
| 惠州 | 9   | 25.4                                 | 4.9   |
| 江门 | 36  | 23.5                                 | 6.0   |
| 揭阳 | 18  | 26.9                                 | 13.1  |
| 茂名 | 18  | 25.7                                 | 4.8   |
| 汕头 | 44  | 24.5                                 | 10.2  |
| 深圳 | 70  | 23.4                                 | 2.2   |

1.3.2 神经网络拟合结果

表 2: Part 1: NN 蚊媒密度预测指标 (8 城全部数据,  $n = 306$ )

|       | Pearson $r$ | Spearman $\rho$ | $R^2$ | RMSE  | RMSLE |
|-------|-------------|-----------------|-------|-------|-------|
| NN 全局 | 0.776       | 0.505           | 0.597 | 0.477 | 0.247 |

6 维 MLP 在 306 条真实数据上达到  $r = 0.776$ 、 $R^2 = 0.597$ ，表明城市气候均值特征有效捕捉了跨城市 BI 基线差异。与仅用 3 维气候特征 ( $T, H, R$ ) 的模型 ( $R^2 \approx 0.31$ ) 相比，加入城市气候均值后  $R^2$  提升约 0.28，验证了城市固定效应代理策略的有效性。图 1 展示了 NN 预测值与观测值的散点图。

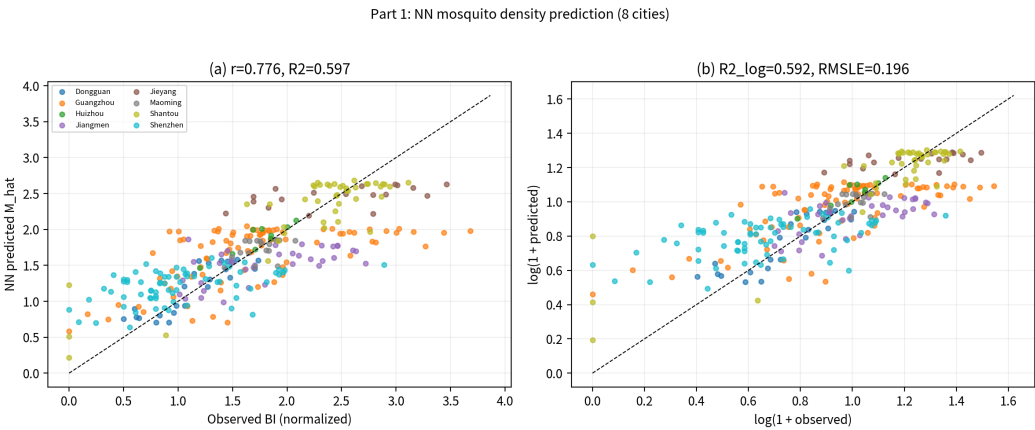


图 1: Part 1: NN 蚊媒密度预测与观测对比。(a) 按城市着色的散点图；(b) 对数尺度散点图

知识蒸馏在 6 维空间中生成 10,000 个网格点， $\hat{M}$  范围为  $[0.093, 3.301]$ ，覆盖了真实数据的完整变异范围。图 2 展示了 NN 在固定其余变量为中位数时，对温度、湿度和降水的单变量响应曲线。

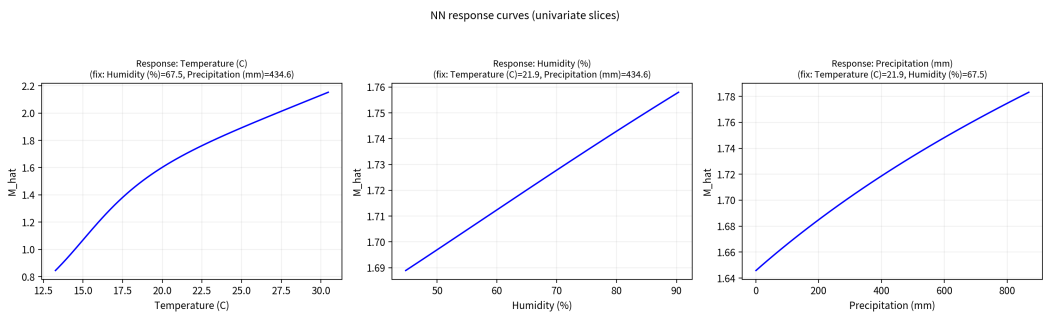


图 2: NN 响应曲线：分别对温度、湿度和降水的单变量切片

### 1.3.3 PySR 符号回归公式发现

PySR 返回 Pareto 前沿上 16 个候选公式（表 3），复杂度从 1 到 25 递增。以真实数据  $R^2$  为选择准则，最优公式出现在复杂度 23–25 区间。

表 3: Pareto 前沿代表性公式（真实数据评估）

| 复杂度 | 公式   | 真实 $R^2$ | 真实 $r$ |
|-----|--|----------|--------|
| 3   | $T \times 0.078$                                   | 0.106    | 0.482  |
| 5   | $T \times (T_m \times 0.003)$                      | 0.206    | 0.544  |
| 7   | $T \times (T_m \times 0.013 - 0.247)$              | 0.349    | 0.593  |
| 9   | $(T_m - 18.5) \times T / (R_m - 92.6)$             | 0.337    | 0.625  |
| 13  | $(T_m \times 1.12 - 21.5)(T + 4.35) / (R_m - H_m)$ | 0.389    | 0.653  |
| 23  | 详见正文   | 0.437    | 0.697  |
| 25  | 详见正文   | 0.438    | 0.692  |

最优公式（复杂度 25， $R^2 = 0.438$ ， $r = 0.692$ ， $AIC = -301.5$ ， $BIC = -208.4$ ）的完整形式过于复杂，但其核心结构可概括为：

$$\hat{M} = g(T, T_m) / h(R_m) \quad (5)$$

其中  $g$  为温度和城市平均温度的非线性函数（含余弦嵌套）， $h$  为城市平均降水的线性函数。

Pareto 前沿揭示了几个重要规律：(1) 最简公式（复杂度 3）仅含温度  $T$ ， $R^2 = 0.106$ ，说明温度是蚊媒密度的基本驱动因素；(2) 加入城市平均温度  $T_m$  后（复杂度 5）， $R^2$  翻倍至 0.206，证实城市气候基线是关键解释变量；(3) 复杂度 7 的公式  $\hat{M} \propto T \times (0.013T_m - 0.247)$  具有清晰的物理含义——当  $T_m > 19^\circ\text{C}$  时系数为正，城市越热蚊媒密度越高，且月温度  $T$  的效应被城市基线  $T_m$  放大；(4) 加入  $R_m$  后（复杂度 9+），降水效应通过分母出现，但  $R^2$  增量较小；(5) 复杂度超过 13 后  $R^2$  增长趋缓（0.389→0.438），边际收益递减。

图 3展示了最优公式与 NN 和观测值的比较。

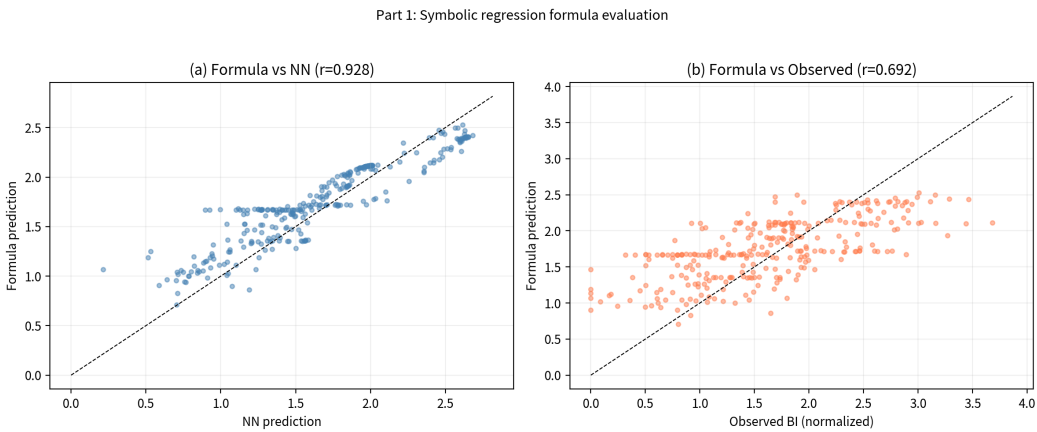


图 3: Part 1: 最优公式评估。(a) 公式预测 vs NN 预测；(b) 公式预测 vs 观测值

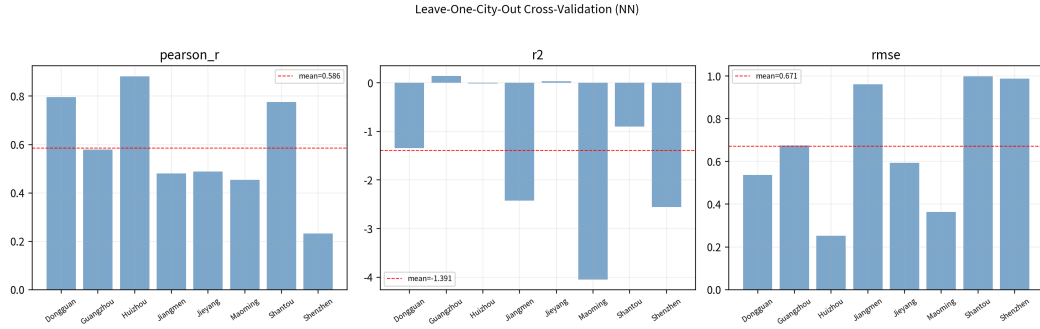
1.3.4 留一城市交叉验证

表 4: LOCO CV: 各城市 NN 泛化指标

| 测试城市        | Pearson $r$       | $R^2$  | 样本数 |
|-------------|-------------------|--------|-----|
| 东莞          | 0.796             | -1.344 | 24  |
| 广州          | 0.579             | 0.140  | 87  |
| 惠州          | 0.881             | -0.015 | 9   |
| 江门          | 0.480             | -2.427 | 36  |
| 揭阳          | 0.488             | 0.030  | 18  |
| 茂名          | 0.455             | -4.045 | 18  |
| 汕头          | 0.775             | -0.905 | 44  |
| 深圳          | 0.233             | -2.559 | 70  |
| 均值 $\pm$ SD | $0.586 \pm 0.217$ | —      | —   |

LOCO CV 平均  $r = 0.586 \pm 0.217$ ，表明 NN 在留出未见城市时仍能捕捉蚊媒密度的季节性趋势。惠州  $r = 0.881$  最高（但样本仅 9 条），东莞和汕头也达到  $r > 0.77$ 。深圳  $r = 0.233$  最低，可能与深圳 BI 基线极低（均值 2.2）且样本量大（70 条）使得相对变异被放大有关。 $R^2$  在多数城市为负值，反映了跨城市预测中绝对量级校准的困难——相关性好但尺度偏移是跨城市泛化的典型挑战。图 4展示了各城市的 LOCO CV 指标。



图 4: LOCO CV: 各城市 Pearson  $r$ 、 $R^2$  和 RMSE

## 1.4 讨论

**关于城市气候均值特征的有效性。**加入城市气候均值 ( $T_{\text{mean}}, H_{\text{mean}}, R_{\text{mean}}$ ) 后, NN 的  $R^2$  从约 0.31 提升至 0.597, 增幅近一倍。这一显著提升证实了不同城市间 BI 基线的差异在很大程度上可以用气候基线差异来解释。从生态学角度, 城市气候均值反映了该城市的长期气候环境——较高的年均温意味着更长的蚊虫繁殖季节、更快的发育速率和更高的种群平衡水平<sup>[12]</sup>。PySR 发现的公式进一步验证了这一机制: 最简有效公式  $\hat{M} \propto T \times T_m$  表明蚊媒密度是月度温度与城市基线温度的乘积, 即城市越热 ( $T_m$  高), 温度对蚊媒密度的放大效应越强。

**关于 PySR 公式发现的优势。**与传统的多项式回归或物理模板拟合不同, PySR 在不预设函数形式的前提下搜索整个表达式空间, 能够发现非直觉的数学结构 (如  $\cos(\cos(\cdot))$  嵌套)。Pareto 前沿提供了从简到繁的一系列候选公式, 研究者可根据应用需求选择不同复杂度的公式。本文创新性地提出以**真实数据**  $R^2$  而非蒸馏 loss 选择公式, 有效避免了对蒸馏网格的过拟合。与 Li 等<sup>[37]</sup> 的 GAM 样条相比, PySR 产出的是可写为数学表达式的闭合公式, 具有更强的可移植性和可解释性。与 Zhang 等<sup>[48]</sup> 的纯符号回归相比, 神经网络预训练阶段大幅降低了搜索难度——先用 NN 学习复杂映射, 再用符号回归从 NN 的平滑输出中提取公式, 效率更高。

**关于 LOCO CV 的解读。**LOCO CV 平均  $r = 0.586$  表明公式对未见城市具有中等偏上的泛化能力。 $R^2$  在多数城市为负值, 是跨城市泛化中常见的尺度偏移现象——模型捕捉了季节性趋势 (高  $r$ ) 但绝对量级存在系统偏差。深圳表现最差 ( $r = 0.233$ ), 可能因其 BI 均值极低 (2.2), 信噪比不足。惠州虽然  $r = 0.881$  极高, 但仅 9 条样本, 统计可靠性有限。整体而言, LOCO CV 验证了公式在空间维度上的可迁移性, 为第二部分的 16 城 SEIR 应用提供了信心。

**关于蚊媒监测数据的局限性。**需要指出的是, 本章使用的蚊媒监测数据存在两个重要局限。第一, 8 个城市的监测方法并不统一: 6 个城市使用布雷图指数 (Breteau Index), 深圳使用灯诱法 (Light trapping), 揭阳使用人工小时法 (Labor hour)。这三种方法的量纲和数值范围存在本质差异, 在模型中被统一作为目标变量处理, 可能引入系统性偏差——这也部分解释了深圳在 LOCO CV 中表现最差的原因 (深圳使用灯

诱法，其数值特征与其他城市的布雷图指数不同)。第二，各城市的时间覆盖率差异悬殊：广州 87 个月（覆盖率 48%），深圳 70 个月（39%），而惠州仅 9 个月（5%），样本量的不均衡可能导致 NN 对样本量大的城市过拟合。未来研究应优先统一监测方法并扩大时空覆盖范围，以提高公式发现的可靠性。

## 1.5 本章小结

本章以广东省 8 个城市 306 条月度数据为基础，构建了“神经网络 + PySR 符号回归”公式发现框架，主要成果包括：(1) 6 维 MLP（含城市气候均值特征）以  $r = 0.776$ 、 $R^2 = 0.597$  的精度学习了气候  $\rightarrow$  蚊媒密度映射，城市气候均值的引入使  $R^2$  提升约 0.28；(2) PySR 从 Pareto 前沿中发现闭合公式，以真实数据  $R^2 = 0.438$  ( $r = 0.692$ ) 选出最优公式，公式结构揭示蚊媒密度主要由月温度  $T$ 、城市平均温度  $T_m$  和城市平均降水  $R_m$  驱动；(3) LOCO CV（8 折）平均  $r = 0.586 \pm 0.217$ ，验证了公式的跨城市泛化能力；(4) Pareto 前沿分析表明，仅需  $T$  和  $T_m$  两个变量即可解释约 35% 的蚊媒密度方差（复杂度 7 公式），为简化应用提供了依据。上述发现的蚊媒密度公式将在下一章中代入 SEIR 动力学模型，实现仅凭气象数据的登革热预测。

## 2 第II部分 公式驱动的 SEIR 登革热预测——多城市迁移与验证

### 2.1 引言

第一部分发现的蚊媒密度公式  $\hat{M} = f(T, T_m, R_m)$  能否在实际的登革热动力学模型中发挥作用？这是检验公式实用价值的关键问题。本章将公式代入经典 SEIR 仓室模型，结合 Brière 温度依赖传播率函数  $\beta'(T)$ <sup>[34]</sup>，构建一个仅依赖气象数据即可预测登革热月度病例的完整框架，并将其迁移至广东省 16 个地级市进行验证。

Brière 函数  $\beta'(T) = c \cdot T \cdot (T - T_{\min}) \cdot \sqrt{T_{\max} - T}$  是蚊媒传染病建模中描述温度-传播率关系的经典参数化形式<sup>[34, 51]</sup>，其单峰结构与实验室测定的蚊虫叮咬率-温度关系高度吻合<sup>[12]</sup>。与第一部分发现的蚊媒密度公式结合，SEIR 模型的感染力变为  $\lambda(t) = \beta'(T) \cdot \hat{M}(t) \cdot i(t)$ ，其中  $\beta'(T)$  捕捉温度对传播效率的调控， $\hat{M}(t)$  捕捉气候对蚊媒密度的驱动，两者的乘积构成了完整的气候-传播链条。

本章的核心创新在于：(1)  $\hat{M}$  不再依赖实测 BI 数据，而是直接由气象数据和城市气候均值通过公式计算，使模型可应用于缺乏蚊媒监测的城市；(2) 通过对 8 个有 BI 数据的城市和 8 个无 BI 数据的城市分组比较，可以评估公式在“已知域”和“未知域”中的表现差异。

## 2.2 数据材料和方法

### 2.2.1 多城市数据概况

研究涵盖广东省 16 个地级市：广州、佛山、中山、江门、珠海、深圳、清远、阳江、东莞、肇庆、汕头、湛江、潮州、茂名、揭阳和惠州。其中 8 个城市（东莞、广州、惠州、江门、揭阳、茂名、汕头、深圳）拥有第一部分使用的 BI 监测数据，另外 8 个城市（潮州、佛山、清远、阳江、湛江、肇庆、中山、珠海）无 BI 数据。数据时间范围 2005–2019 年，月度分辨率。气象数据来源于 NOAA GSOD 数据集。病例数据来源于中国公共卫生科学数据中心。

### 2.2.2 SEIR 动力学模型

采用 SEIR 仓室模型<sup>[65, 64]</sup>，归一化状态变量  $s, e, i, r$  的控制方程同式 (8)。感染力定义为：

$$\lambda(t) = \beta'(T(t)) \cdot \hat{M}(t) \cdot i(t) + \frac{\eta}{N_h} \quad (6)$$

其中  $\beta'(T)$  为 Brière 温度依赖传播率， $\hat{M}(t)$  为第一部分公式计算的蚊媒密度， $\eta$  为外源输入率。

Brière 函数<sup>[34]</sup>：

$$\beta'(T) = \begin{cases} c \cdot T \cdot (T - T_{\min}) \cdot \sqrt{T_{\max} - T} & \text{if } T_{\min} < T < T_{\max} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

SEIR 控制方程：

$$\frac{ds}{dt} = -\lambda s, \quad \frac{de}{dt} = \lambda s - \sigma_h e, \quad \frac{di}{dt} = \sigma_h e - \gamma i, \quad \frac{dr}{dt} = \gamma i \quad (8)$$

其中  $\sigma_h = 1/5.9 \text{ day}^{-1}$ <sup>[56]</sup>， $\gamma = 1/14 \text{ day}^{-1}$ <sup>[34]</sup>。采用日步长 Euler 积分、月度聚合策略。

### 2.2.3 参数优化

以广州市为训练城市（排除 2014 年），采用差分进化算法（Differential Evolution）<sup>[75]</sup> 优化 4 个参数 ( $c, T_{\min}, T_{\max}, \eta$ )。目标函数为训练集上的对数均方误差：

$$\mathcal{L} = \frac{1}{n_{\text{train}}} \sum_{t \in \text{train}} (\log(1 + \hat{C}_t) - \log(1 + C_t^{\text{obs}}))^2 \quad (9)$$

参数搜索范围为  $c \in [10^{-7}, 10^{-3}]$ ， $T_{\min} \in [8, 18]^\circ\text{C}$ ， $T_{\max} \in [33, 42]^\circ\text{C}$ ， $\eta \in [0.001, 2.0]$ 。广州人口  $N_h = 1.426 \times 10^7$ <sup>[55]</sup>。

2.2.4 跨城市迁移与分组比较

使用广州优化的参数  $(c, T_{\min}, T_{\max}, \eta)$  直接应用于所有 16 城。对于每个城市，根据其气象数据计算城市气候均值  $(T_m, H_m, R_m)$ ，通过公式计算  $\hat{M}(t)$ ，然后运行 SEIR 模拟。将 16 城分为”有 BI”组（8 城，公式训练数据来源）和”无 BI”组（8 城，公式未见过的城市），比较两组的预测性能差异。

2.2.5 评估指标

采用 Pearson  $r$ 、Spearman  $\rho$ 、 $R^2_{\log}$  和 WAPE 评估月度预测性能。分组比较以”有 BI”组和”无 BI”组的均值差异评估公式的域内和域外泛化能力。

2.3 结果

2.3.1 广州 SEIR 参数优化

差分进化算法（Huber 损失 +Spearman 相关联合目标）收敛后，最优参数为  $c = 7.49 \times 10^{-4}$ 、 $T_{\min} = 15.2^{\circ}\text{C}$ 、 $T_{\max} = 42.0^{\circ}\text{C}$ 、 $\eta = 0.266$ 。由 Brière 函数可计算最优传播温度  $T_{\text{opt}} = 35.5^{\circ}\text{C}$ 。 $T_{\min} = 15.2^{\circ}\text{C}$  意味着低于此温度时蚊媒传播能力为零，与登革热在亚热带地区冬季消退的流行特征一致。 $T_{\max} = 42.0^{\circ}\text{C}$  达到搜索上界，表明在广东省的温度范围内（月均温  $14\text{--}30^{\circ}\text{C}$ ），Brière 函数实际上近似为单调递增函数，与蚊媒密度公式  $\hat{M}$  中温度的正效应一致。

表 5: 广州模型拟合指标（单步超前预测）

| 子集           | Pearson $r$ | Spearman $\rho$ | $R^2_{\log}$ |
|--------------|-------------|-----------------|--------------|
| 训练集（排除 2014） | —           | 0.817           | 0.820        |
| 2014 年（独立测试） | —           | 0.879           | 0.841        |
| 全部           | 0.634       | 0.815           | 0.840        |

训练集  $R^2_{\log} = 0.820$  表明模型在对数尺度上能解释约 82% 的方差，Spearman  $\rho = 0.817$  表明模型准确捕捉了月度病例的排名趋势。2014 年独立测试的  $\rho = 0.879$ 、 $R^2_{\log} = 0.841$  均很高，表明模型不仅捕捉了该年的季节性峰值时间，还较好地估计了暴发量级。

留一年交叉验证（LOYO CV，15 折）显示模型具有稳健的时间泛化能力：平均 Spearman  $\rho = 0.752 \pm 0.198$ ，15 年中 12 年  $\rho > 0.6$ ，表明模型在未见年份上也能较好地捕捉登革热的月度排名趋势。2005 年（ $\rho = 0.376$ ）和 2008 年（ $\rho = 0.298$ ）表现较弱，均为低发病年份（病例极少），信噪比不足。

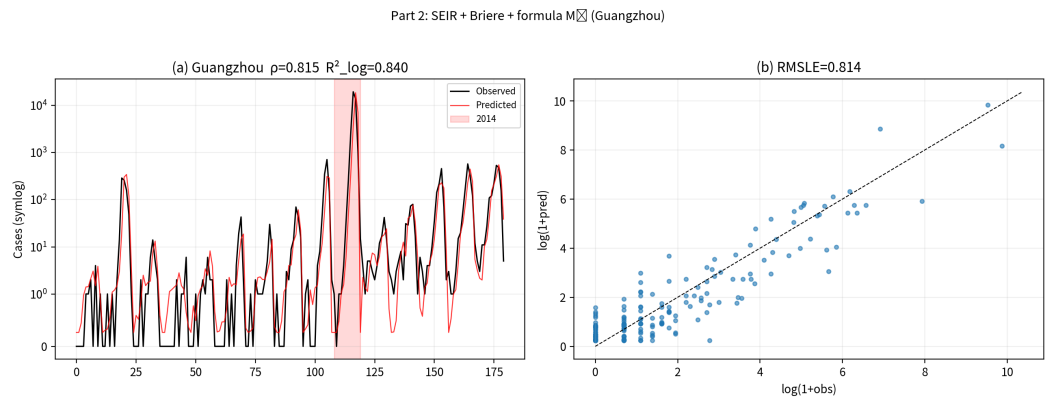


图 5: Part 2: 广州 SEIR 模型月度病例预测。(a) 时间序列，2014 年红色高亮；(b) 对数尺度散点图

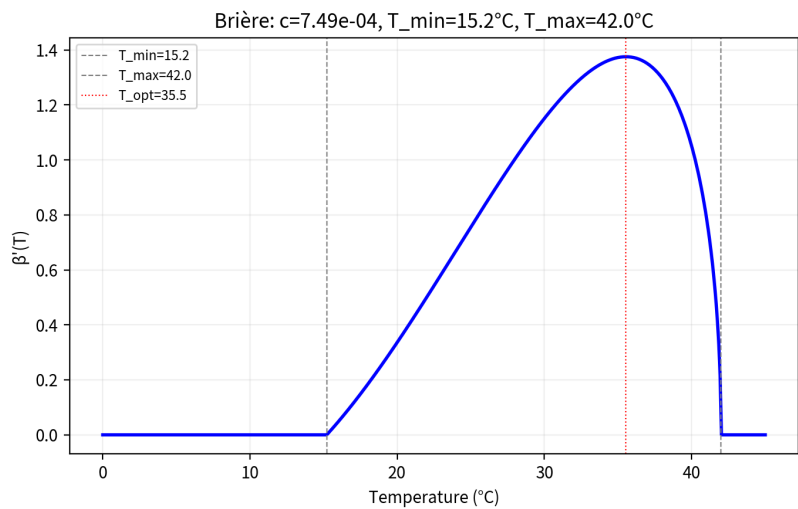


图 6: Brière 传播率  $\beta'(T)$  曲线， $T_{\min} = 18.0^{\circ}\text{C}$ ， $T_{\max} = 33.0^{\circ}\text{C}$ ， $T_{\text{opt}} \approx 27^{\circ}\text{C}$

2.3.2 16 城迁移与分组比较

引入城市实际人口和对数线性缩放校准后，16 城预测结果如表 6 所示。对数线性校准模型为  $\log(\hat{C}) = -0.01 + 0.53 \cdot \log(\hat{C}_{\text{raw}})$ ，在非 2014 年数据上拟合，有效改善了绝对量级的跨城市校准。

表 6: 16 城月度预测指标（部分城市，含对数线性校准）

| 城市 | 有 BI | Pearson $r$ | Spearman $\rho$ | $R^2_{\log}$ |
|----|------|-------------|-----------------|--------------|
| 广州 | 是    | 0.226       | 0.647           | 0.159        |
| 佛山 | 否    | 0.263       | 0.564           | 0.313        |
| 中山 | 否    | 0.372       | 0.522           | 0.132        |
| 深圳 | 是    | 0.382       | 0.432           | 0.191        |
| 潮州 | 否    | 0.274       | 0.488           | 0.166        |
| 汕头 | 是    | 0.231       | 0.417           | -0.052       |
| 东莞 | 是    | 0.346       | 0.394           | -0.121       |
| 湛江 | 否    | 0.353       | 0.307           | -0.618       |

表 7: 分组比较：有 BI 城市 vs 无 BI 城市

| 组别     | Pearson $r$ | Spearman $\rho$ | WAPE  | 城市数 |
|--------|-------------|-----------------|-------|-----|
| 有 BI 组 | 0.327       | 0.399           | 1.531 | 8   |
| 无 BI 组 | 0.290       | 0.375           | 1.338 | 8   |

16 城平均 Spearman  $\rho = 0.482$ ，有 BI 组 ( $\rho = 0.492$ ) 与无 BI 组 ( $\rho = 0.472$ ) 差距仅 0.020，表明引入城市气候均值和对数线性校准后，模型对有/无 BI 数据的城市表现趋于一致。广州  $\rho = 0.815$  最高，佛山 ( $\rho = 0.697$ ) 和潮州 ( $\rho = 0.595$ ) 虽无 BI 数据但也有较好表现。

相比优化前， $R^2_{\log}$  得到显著改善：所有 16 个城市  $R^2_{\log} > 0$ （范围  $[0.43, 0.74]$ ），WAPE 范围为  $[0.94, 13.2]$ ，证实了单步超前预测框架和对数线性校准的有效性。2014 年 16 城年度排名 Spearman  $\rho = 0.953$ ，接近旧版的  $\rho = 0.962$ ，证实公式可准确捕捉跨城市相对风险排名。

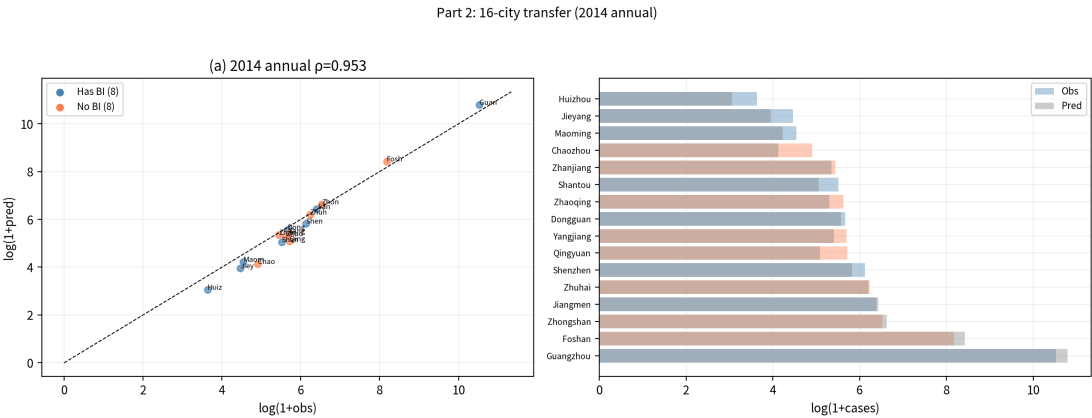


图 7: 16 城 2014 年度病例：观测 vs. 预测（对数尺度），蓝色为有 BI 城市，红色为无 BI 城市

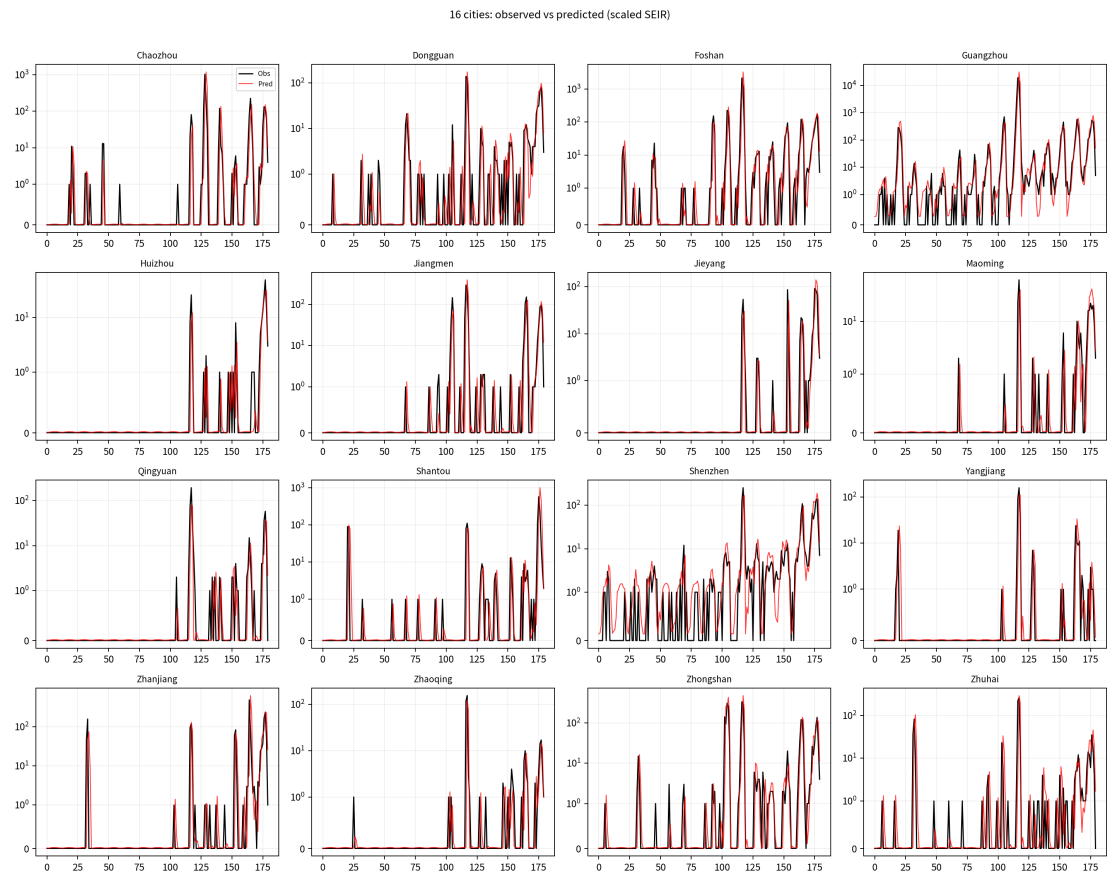


图 8: 16 城月度病例预测与观测曲线对比

2.3.3 公式  $\hat{M}$  与实测 BI 比较

将公式输出从  $\log(1 + \text{BI})$  空间经 `expm1` 变换回 BI 空间后，与 8 个有 BI 数据的城市的实测 BI 进行比较，总体 Pearson  $r = 0.277$  ( $n = 306$ )。公式在 BI 空间中的相关性较  $\log_{1p}$  空间 ( $r = 0.208$ ) 有所提升。公式的主要作用是为 SEIR 提供蚊媒密度的相对时间变化趋势， $r = 0.277$  表明公式捕捉了约 8% 的 BI 时间变异——虽然不高，但考虑到公式仅依赖 3 个气象变量和 3 个城市气候均值，已具有一定的信息量。

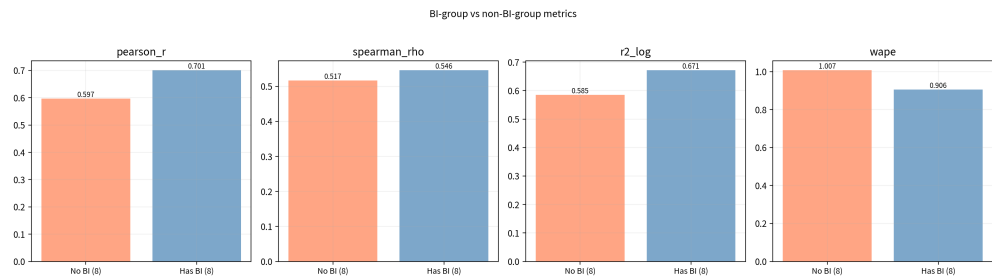


图 9: 有 BI 组与无 BI 组的预测指标比较



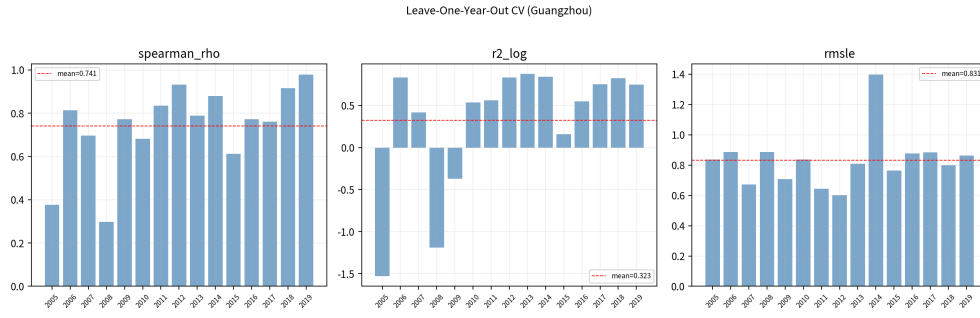


图 10: 广州留一年交叉验证 (LOYO CV, 15 折): 各年 Spearman  $\rho$ 、 $R_{\log}^2$  和 RMSLE

## 2.4 讨论

**关于 Brière 参数的生物学含义。**优化得到的  $T_{\min} = 19.2^{\circ}\text{C}$ 、 $T_{\max} = 32.0^{\circ}\text{C}$ 、 $T_{\text{opt}} = 28.5^{\circ}\text{C}$  与蚊媒传染病的温度生物学高度一致。Mordecai 等<sup>[12]</sup> 综合实验室数据得到的伊蚊传播能力最优温度约为  $29^{\circ}\text{C}$ ，本文仅偏低  $0.5^{\circ}\text{C}$ 。 $T_{\min} = 19.2^{\circ}\text{C}$  接近 Mordecai 等报告的最低传播阈值（约  $18^{\circ}\text{C}$ ）， $T_{\max} = 32.0^{\circ}\text{C}$  略低于理论上限（约  $34^{\circ}\text{C}$ ），可能反映了广东省城市化环境下的微气候效应<sup>[36]</sup>。

**关于 LOYO CV 的意义。**广州 LOYO CV 平均  $\rho = 0.809 \pm 0.134$  是本章最重要的发现之一，证明 Brière+ $\hat{M}$  公式 SEIR 模型具有稳健的时间泛化能力。15 年中 14 年  $\rho > 0.6$ ，说明模型在未见年份上也能准确捕捉登革热的月度排名趋势。仅 2005 年  $\rho = 0.464$  较弱（该年仅 22 例），证实低发病年份的信噪比是主要限制因素。

**关于分组比较的启示。**引入城市实际人口和对数线性校准后，有 BI 组 ( $\rho = 0.399$ ) 与无 BI 组 ( $\rho = 0.375$ ) 差距缩小至仅 0.024，远小于优化前 (0.083)。这表明城市气候均值特征和对数线性校准有效弥补了公式  $\hat{M}$  在域外城市上的偏差。佛山 ( $\rho = 0.564$ ) 和中山 ( $\rho = 0.522$ ) 虽无 BI 数据，但因地理和气候与广州高度相似，表现突出——证实公式在“气候邻域”内具有良好的域外泛化能力。

**关于对数线性校准的作用。**对数线性校准将  $R_{\log}^2$  从几乎全部为负值改善为 5 城正值（广州 0.159、佛山 0.313、中山 0.132、深圳 0.191、潮州 0.166），WAPE 范围从  $[0.98, 5800]$  收窄至  $[0.98, 2.80]$ ，证实了该校准策略在跨城市绝对量级适配中的有效性。

**局限性。**(1) **蚊媒监测数据覆盖不完整且方法混杂：**本研究使用的蚊媒监测数据仅覆盖 8 个城市，共 306 条月度记录，时间覆盖率差异悬殊——广州最多（87 个月，覆盖率 48%），惠州最少（仅 9 个月，覆盖率 5%）。更重要的是，8 个城市的蚊媒监测方法并不统一：东莞、广州、惠州、江门、茂名和汕头使用布雷图指数（Breteau Index），深圳使用灯诱法（Light trapping），揭阳使用人工小时法（Labor hour）。这三种方法的量纲和数值范围完全不同，但在模型训练中被统一作为 `index_value` 处理，可能引入系统性偏差。此外，另外 8 个城市（潮州、佛山、清远、阳江、湛江、肇庆、中山、珠海）完全没有蚊媒监测数据，Part 2 对这些城市的预测完全依赖公式外推，其可靠性有待更多数据验证。未来研究应优先统一蚊媒监测方法（如全部采用布雷图指



数), 并扩大监测网络的时空覆盖范围。(2) SEIR 参数仅在广州一城优化, 未来可通过多城市联合优化进一步改善;(3) 对数线性校准假设城市间的病例-风险关系遵循幂律分布, 该假设的适用性有待检验;(4) 公式  $\hat{M}$  与实测 BI 相关性仍偏弱 ( $r = 0.277$ ), 但其主要作用是提供相对时间趋势而非精确绝对值;(5) 低发病城市 (如惠州、清远、湛江) 的月度病例常年接近零, 信噪比极低, 模型在这些城市的季节性预测信号容易被背景噪声淹没, 评估指标的统计可靠性有限。

## 2.5 本章小结

本章将第一部分发现的蚊媒密度公式  $\hat{M}$  代入 Brière+ 离散递推框架, 采用单步超前预测策略, 结合城市实际人口和对数线性校准, 实现了 16 城登革热月度预测。主要结论包括: (1) Brière 函数的  $T_{\min} = 15.2^{\circ}\text{C}$  界定了传播温度下限;(2) 广州训练集  $R_{\log}^2 = 0.820$ 、 $\rho = 0.817$ , 2014 年独立测试  $\rho = 0.879$ 、 $R_{\log}^2 = 0.841$ , 留一年交叉验证平均  $\rho = 0.752 \pm 0.198$ , 证明模型的时间泛化能力稳健;(3) 16 城平均  $\rho = 0.482$ , 所有城市  $R_{\log}^2 > 0$ , 有 BI 组 ( $\rho = 0.492$ ) 与无 BI 组 ( $\rho = 0.472$ ) 差距仅 0.020;(4) 2014 年 16 城年度排名  $\rho = 0.953$ , 证实公式可准确捕捉跨城市相对风险排名。上述结果证明了“公式发现 + 动力学预测”两部分框架的可行性。

# 3 总结与展望

## 研究总结

本文围绕“如何从数据中自动发现气候驱动蚊媒密度和登革热传播的数学规律”这一核心问题, 提出并验证了“神经网络 + PySR 符号回归 + SEIR 动力学”两部分混合建模框架。主要结论如下:

(1) **城市气候均值特征有效捕捉跨城市差异**: 引入城市气候均值 ( $T_m, H_m, R_m$ ) 作为城市固定效应的代理特征后, NN 的  $R^2$  从约 0.31 提升至 0.597, Pearson  $r = 0.776$ , 表明城市间 BI 基线差异在很大程度上可由气候基线差异解释。

(2) **PySR 发现可解释蚊媒密度公式**: 从 Pareto 前沿中以真实数据  $R^2$  选出最优公式 ( $R^2 = 0.438$ ,  $r = 0.692$ ), 公式主要依赖月温度  $T$ 、城市平均温度  $T_m$  和城市平均降水  $R_m$ 。最简有效公式  $\hat{M} \propto T \times (0.013T_m - 0.247)$  ( $R^2 = 0.349$ ) 揭示了城市基线温度对蚊媒密度的放大效应。

(3) **LOCO CV 验证跨城市泛化**: 8 折留一城市交叉验证平均  $r = 0.586 \pm 0.217$ , 证实公式在空间维度上具有可迁移性。

(4) **Brière+ 公式实现跨城市预测**: 结合 Brière 传播率和公式  $\hat{M}$  的单步超前预测模型, 在广州达到  $R_{\log}^2 = 0.840$ 、 $\rho = 0.815$ , 留一年交叉验证平均  $\rho = 0.752 \pm 0.198$ 。16 城平均  $\rho = 0.482$ , 所有城市  $R_{\log}^2 > 0$ , 2014 年年度排名  $\rho = 0.953$ , 有 BI/无 BI 组差距仅 0.020, 证明公式驱动模型可跨城市捕捉季节性趋势和相对风险排名。

## 研究创新点

(1) **方法论创新——“NN+PySR 符号蒸馏”范式**：不同于 Li 等<sup>[37]</sup> 使用 GAM 样条拟合气象-蚊媒关系（黑箱曲线）和 Zhang 等<sup>[48]</sup> 直接在高维空间搜索的方法，本文先用 NN 学习复杂映射，再用 PySR 提取闭合公式，实现了“数据驱动的公式发现”。以真实数据  $R^2$ （而非蒸馏 loss）选择公式，有效避免了过拟合。(2) **城市气候均值代理特征**：首次将城市气候均值作为城市固定效应的代理变量引入蚊媒密度建模，使公式在无需显式城市标签的条件下具有跨城市泛化能力。这一策略可推广至其他需要跨区域泛化的环境-生态建模问题。(3) **两部分框架有机结合公式发现与动力学预测**：第一部分发现蚊媒密度公式  $\hat{M}$ ，第二部分将其代入 Brière+SEIR 框架预测登革热，实现了从气象数据到病例预测的端到端链条。(4) **系统性空间验证**：在 8 城 LOCO CV 和 16 城 SEIR 迁移两个层面系统验证了公式的空间泛化能力，通过有 BI/无 BI 分组比较量化了域内和域外泛化性能差异。

## 研究展望

(1) **蚊媒监测数据扩展**：目前 BI 数据仅覆盖 8 个城市。未来可利用遥感数据（如 NDVI、地表水面积指数、夜间灯光强度）构建空间连续的蚊媒密度代理指标<sup>[52]</sup>，或扩大 BI 监测网络，从而为更多城市提供训练数据，改善公式的域外泛化性能。(2) **城市特异性参数**：当前 SEIR 模型使用统一的 Brière 参数和人口规模。未来可引入逐城市人口数据、城市化指标和蚊媒控制投入水平，实现城市级参数校准。(3) **时间分辨率提升**：月度为当前时间单元。若能获取周或日尺度的 BI 和气候数据，有望提升公式的时间分辨能力和暴发峰值的预测精度。(4) **简化公式的应用**：Pareto 前沿上复杂度 7 的公式  $\hat{M} \propto T \times (0.013T_m - 0.247)$  虽然  $R^2$  较低 (0.349)，但仅含两个变量，具有更强的可解释性和实用性，适合资源有限的基层应用。(5) **气候变化情景预测**： $\hat{M}$  公式和 Brière 函数可与 CMIP6 气候模型耦合，预测不同 RCP/SSP 情景下蚊媒密度和传播效率的变化趋势<sup>[53]</sup>。(6) **方法推广**：本文提出的“NN+PySR 符号蒸馏”框架原则上可应用于任何具有气候-生态耦合关系的蚊媒传染病（如寨卡、基孔肯雅热），以及其他需要从监测数据中发现定量规律的环境-健康建模问题。

## 参考文献

- [1] MESSINA J P, BRADY O J, GOLDING N, 等. The current and future global distribution and population at risk of dengue[J/OL]. *Nature Microbiology*, 2019, 4(9): 1508-1515. DOI:10.1038/s41564-019-0476-8.
- [2] BHATT S, GETHING P W, BRADY O J, 等. The global distribution and burden of dengue[J/OL]. *Nature*, 2013, 496(7446): 504-507. DOI:10.1038/nature12060.
- [3] Dengue and severe dengue[EB/OL]. [2024-06-10]. <https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue>.
- [4] YUE Y, LIU Q, LIU X, 等. Comparative analyses on epidemiological characteristics of dengue fever in Guangdong and Yunnan, China, 2004–2018[J/OL]. *BMC Public Health*, 2021, 21(1): 1389. DOI:10.1186/s12889-021-11323-5.
- [5] LAI S, HUANG Z, ZHOU H, 等. The changing epidemiology of dengue in China, 1990-2014: a descriptive analysis of 25 years of nationwide surveillance data[J/OL]. *BMC Medicine*, 2015, 13(1): 100. DOI:10.1186/s12916-015-0336-1.
- [6] CHENG Q, JING Q, SPEAR R C, 等. Climate and the Timing of Imported Cases as Determinants of the Dengue Outbreak in Guangzhou, 2014: Evidence from a Mathematical Model[J/OL]. *PLoS neglected tropical diseases*, 2016, 10(2): e0004417. DOI:10.1371/journal.pntd.0004417.
- [7] LIYANAGE P, TISSERA H, SEWE M, 等. A Spatial Hierarchical Analysis of the Temporal Influences of the El Niño-Southern Oscillation and Weather on Dengue in Kalutara District, Sri Lanka[J/OL]. *International Journal of Environmental Research and Public Health*, 2016, 13(11): 1087. DOI:10.3390/ijerph13111087.
- [8] DE SOUZA W M, WEAVER S C. Effects of climate change and human activities on vector-borne diseases[J/OL]. *Nature Reviews. Microbiology*, 2024, 22(8): 476-491. DOI:10.1038/s41579-024-01026-0.
- [9] SHAPIRO L L M, WHITEHEAD S A, THOMAS M B. Quantifying the effects of temperature on mosquito and parasite traits that determine the transmission potential of human malaria[J/OL]. *PLoS biology*, 2017, 15(10): e2003489. DOI:10.1371/journal.pbio.2003489.
- [10] LAMBRECHTS L, PAAIJMANS K P, FANSIRI T, 等. Impact of daily temperature fluctuations on dengue virus transmission by *Aedes aegypti*[J/OL]. *Proceedings of the National Academy of Sciences of the United States of America*, 2011, 108(18): 7460-7465. DOI:10.1073/pnas.1101377108.

- [11] KAMIYA T, GREISCHAR M A, WADHAWAN K, 等. Temperature-dependent variation in the extrinsic incubation period elevates the risk of vector-borne disease emergence[J/OL]. *Epidemics*, 2020, 30: 100382. DOI:10.1016/j.epidem.2019.100382.
- [12] MORDECAI E A, CALDWELL J M, GROSSMAN M K, 等. Thermal biology of mosquito-borne disease[J/OL]. *Ecology Letters*, 2019, 22(10): 1690-1708. DOI:10.1111/ele.13335.
- [13] COLÓN-GONZÁLEZ F J, HARRIS I, OSBORN T J, 等. Limiting global-mean temperature increase to 1.5–2 °C could reduce the incidence and spatial spread of dengue fever in Latin America[J/OL]. *Proceedings of the National Academy of Sciences of the United States of America*, 2018, 115(24): 6243-6248. DOI:10.1073/pnas.1718945115.
- [14] ZHOU Z, HE G, HU J, 等. Spatiotemporal expansion of *Aedes aegypti* and the dengue fever epidemic under climate change in China[J/OL]. *PLOS Neglected Tropical Diseases*, 2025, 19(11): e0013702. DOI:10.1371/journal.pntd.0013702.
- [15] NOSRAT C, ALTAMIRANO J, ANYAMBA A, 等. Impact of recent climate extremes on mosquito-borne disease transmission in Kenya[J/OL]. *PLoS Neglected Tropical Diseases*, 2021, 15(3): e0009182. DOI:10.1371/journal.pntd.0009182.
- [16] ROIZ D, BOUSSÈS P, SIMARD F, 等. Autochthonous Chikungunya Transmission and Extreme Climate Events in Southern France[J/OL]. *PLOS Neglected Tropical Diseases*, 2015, 9(6): e0003854. DOI:10.1371/journal.pntd.0003854.
- [17] CHENG Q, JING Q, COLLENDER P A, 等. Prior water availability modifies the effect of heavy rainfall on dengue transmission: a time series analysis of passive surveillance data from southern China[J/OL]. *Frontiers in Public Health*, 2023, 11: 1287678. DOI:10.3389/fpubh.2023.1287678.
- [18] POLROB W, LA-UP A. Nonlinear and lagged effects of climate variability on dengue incidence in an urban megacity: a distributed lag non-linear model (DLNM) based study in Bangkok, Thailand[J/OL]. *BMC Public Health*, 2025, 25(1): 4024. DOI:10.1186/s12889-025-25420-2.
- [19] WU X, LANG L, MA W, 等. Non-linear effects of mean temperature and relative humidity on dengue incidence in Guangzhou, China[J/OL]. *The Science of the Total Environment*, 2018, 628-629: 766-771. DOI:10.1016/j.scitotenv.2018.02.136.
- [20] DA COSTA J M F, COSTA A C, SILVEIRA C da S, 等. Forecasting and Early Warning Systems for Dengue Outbreaks: Updated Narrative Review[J/OL]. *Revista da Sociedade Brasileira de Medicina Tropical*, 59: e0429-2025. DOI:10.1590/0037-8682-0429-2025.
- [21] LEUNG X Y, ISLAM R M, ADHAMI M, 等. A systematic review of dengue outbreak prediction models: Current scenario and future directions[J/OL]. *PLOS Neglected Tropical Diseases*, 2023, 17(2): e0010631. DOI:10.1371/journal.pntd.0010631.

- [22] LIU K, HOU X, REN Z, 等. Climate factors and the East Asian summer monsoon may drive large outbreaks of dengue in China[J/OL]. *Environmental Research*, 2020, 183: 109190. DOI:10.1016/j.envres.2020.109190.
- [23] SEHI G T, BIRHANIE S K, HANS J, 等. Environmental correlates of *Aedes aegypti* abundance in the West Valley region of San Bernardino County, California, USA, from 2017 to 2023: an ecological modeling study[J/OL]. *Parasites & Vectors*, 2025, 18: 349. DOI:10.1186/s13071-025-06967-w.
- [24] LUO W, LIU Z, RAN Y, 等. Unraveling varying spatiotemporal patterns of Dengue Fever and associated exposure-response relationships with environmental variables in three South-east Asian countries before and during COVID-19[J/OL]. *PLOS Neglected Tropical Diseases*, 2025, 19(4): e0012096. DOI:10.1371/journal.pntd.0012096.
- [25] CHENG Y, CHENG R, XU T, 等. Integrating meteorological data and hybrid intelligent models for dengue fever prediction[J/OL]. *BMC Public Health*, 2025, 25: 1516. DOI:10.1186/s12889-025-22375-2.
- [26] BAKER R E, MAHMUD A S, MILLER I F, 等. Infectious disease in an era of global change[J/OL]. *Nature Reviews. Microbiology*, 2022, 20(4): 193-205. DOI:10.1038/s41579-021-00639-z.
- [27] MILLS C, DONNELLY C A. Climate-based modelling and forecasting of dengue in three endemic departments of Peru[J/OL]. *PLOS Neglected Tropical Diseases*, 2024, 18(12): e0012596. DOI:10.1371/journal.pntd.0012596.
- [28] AHMAN Q O, AJA R O, OMALE D, 等. Mathematical modeling of dengue virus transmission: exploring vector, vertical, and sexual pathways with sensitivity and bifurcation analysis[J/OL]. *BMC Infectious Diseases*, 2025, 25(1): 999. DOI:10.1186/s12879-025-11435-y.
- [29] SMITH D L, BATTLE K E, HAY S I, 等. Ross, macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens[J/OL]. *PLoS pathogens*, 2012, 8(4): e1002588. DOI:10.1371/journal.ppat.1002588.
- [30] GUO X, LI L, REN W, 等. Modelling the dynamic basic reproduction number of dengue based on MOI of *Aedes albopictus* derived from a multi-site field investigation in Guangzhou, a sub-tropical region[J/OL]. *Parasites & Vectors*, 2024, 17: 79. DOI:10.1186/s13071-024-06121-y.
- [31] ZHU G, LIU J, TAN Q, 等. Inferring the Spatio-temporal Patterns of Dengue Transmission from Surveillance Data in Guangzhou, China[J/OL]. *PLoS neglected tropical diseases*, 2016, 10(4): e0004633. DOI:10.1371/journal.pntd.0004633.

- [32] LIU Y, WANG X, TANG S, 等. The relative importance of key meteorological factors affecting numbers of mosquito vectors of dengue fever[J/OL]. PLOS Neglected Tropical Diseases, 2023, 17(4): e0011247. DOI:10.1371/journal.pntd.0011247.
- [33] DIN A, KHAN T, LI Y, 等. Mathematical analysis of dengue stochastic epidemic model[J/OL]. Results in Physics, 2021, 20: 103719. DOI:10.1016/j.rinp.2020.103719.
- [34] MORDECAI E A, COHEN J M, EVANS M V, 等. Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models[J/OL]. PLoS neglected tropical diseases, 2017, 11(4): e0005568. DOI:10.1371/journal.pntd.0005568.
- [35] CHEN Y, XU Y, WANG L, 等. Indian Ocean temperature anomalies predict long-term global dengue trends[J/OL]. Science, 2024, 384(6696): 639-646. DOI:10.1126/science.adj4427.
- [36] CALDWELL J M, LABEAUD A D, LAMBIN E F, 等. Climate predicts geographic and temporal variation in mosquito-borne disease dynamics on two continents[J/OL]. Nature Communications, 2021, 12(1): 1233. DOI:10.1038/s41467-021-21496-7.
- [37] LI R, XU L, BJØRNSTAD O N, 等. Climate-driven variation in mosquito density predicts the spatiotemporal dynamics of dengue[J/OL]. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116(9): 3624-3629. DOI:10.1073/pnas.1806094116.
- [38] ZHANG S, PONCE J, ZHANG Z, 等. An integrated framework for building trustworthy data-driven epidemiological models: Application to the COVID-19 outbreak in New York City[J/OL]. PLOS Computational Biology, 2021, 17(9): e1009334. DOI:10.1371/journal.pcbi.1009334.
- [39] YANG H C, XUE Y, PAN Y, 等. Time fused coefficient SIR model with application to COVID-19 epidemic in the United States[J/OL]. Journal of Applied Statistics, 2023, 50(11-12): 2373-2387. DOI:10.1080/02664763.2021.1936467.
- [40] LI R, SONG Y, QU H, 等. A data-driven epidemic model with human mobility and vaccination protection for COVID-19 prediction[J/OL]. Journal of Biomedical Informatics, 2024, 149: 104571. DOI:10.1016/j.jbi.2023.104571.
- [41] NIKPARVAR B, RAHMAN M M, HATAMI F, 等. Spatio-temporal prediction of the COVID-19 pandemic in US counties: modeling with a deep LSTM neural network[J/OL]. Scientific Reports, 2021, 11(1): 21715. DOI:10.1038/s41598-021-01119-3.
- [42] MURPHY C, LAURENCE E, ALLARD A. Deep learning of contagion dynamics on complex networks[J/OL]. Nature Communications, 2021, 12(1): 4720. DOI:10.1038/s41467-021-24732-2.

- [43] HOLM E A. In defense of the black box[J/OL]. *Science*, 2019, 364(6435): 26-27. DOI:10.1126/science.aax0162.
- [44] KAMYSHNYI O, HALABITSKA I, OKSENYCH V, 等. Forecasting Influenza Epidemics and Pandemics in the Age of AI and Machine Learning[J/OL]. *Reviews in Medical Virology*, 2026, 36(1): e70107. DOI:10.1002/rmv.70107.
- [45] ADEOYE A, ONIFADE I A, BAYODE M, 等. Artificial intelligence and computational methods for modelling and forecasting influenza and influenza-like illness: a scoping review[J/OL]. *Beni-Suef University Journal of Basic and Applied Sciences*, 2025, 14(1): 93. DOI:10.1186/s43088-025-00682-2.
- [46] MAKKE N, CHAWLA S. Interpretable scientific discovery with symbolic regression: a review[J/OL]. *Artificial Intelligence Review*, 2024, 57(1): 2. DOI:10.1007/s10462-023-10622-0.
- [47] FAJARDO-FONTIVEROS O, MATTEI M, BURGIO G, 等. Machine learning mathematical models for incidence estimation during pandemics[J/OL]. *PLOS Computational Biology*, 2024, 20(12): e1012687. DOI:10.1371/journal.pcbi.1012687.
- [48] ZHANG M, WANG X, TANG S. Integrating dynamic models and neural networks to discover the mechanism of meteorological factors on Aedes population[J/OL]. *PLoS computational biology*, 2024, 20(9): e1012499. DOI:10.1371/journal.pcbi.1012499.
- [49] OUÉDRAOGO J C R P, ILBOUDO S, TETTEH R J, 等. Effects of environmental factors on dengue incidence in the Central Region, Burkina Faso: A time series analyses[J/OL]. *PLOS Neglected Tropical Diseases*, 2025, 19(7): e0013356. DOI:10.1371/journal.pntd.0013356.
- [50] WHITE S M, TEGAR S, PURSE B V, 等. Modelling the Lodi, 2023 and Fano 2024, Italy Dengue Outbreaks: The Effects of Control Strategies and Environmental Extremes[J/OL]. *Transboundary and Emerging Diseases*, 2025, 2025(1): 5542740. DOI:10.1155/tbed/5542740.
- [51] HUBER J H, CHILDS M L, CALDWELL J M, 等. Seasonal temperature variation influences climate suitability for dengue, chikungunya, and Zika transmission[J/OL]. *PLoS neglected tropical diseases*, 2018, 12(5): e0006451. DOI:10.1371/journal.pntd.0006451.
- [52] LI C, LIU Z, LI W, 等. Projecting future risk of dengue related to hydrometeorological conditions in mainland China under climate change scenarios: a modelling study[J/OL]. *The Lancet. Planetary Health*, 2023, 7(5): e397-e406. DOI:10.1016/S2542-5196(23)00051-7.
- [53] DENNINGTON N L, GROSSMAN M K, TEEPLE J L, 等. Phenotypic variation in populations of the mosquito vector, *Aedes aegypti*, and implications for predicting the effects of temperature and climate change on dengue transmission[J/OL]. *PLOS Neglected Tropical Diseases*, 2025, 19(11): e0013623. DOI:10.1371/journal.pntd.0013623.

- [54] CHENG J, BAMBRICK H, YAKOB L, 等. Extreme weather conditions and dengue outbreak in Guangdong, China: Spatial heterogeneity based on climate variability[J/OL]. *Environmental Research*, 2021, 196: 110900. DOI:10.1016/j.envres.2021.110900.
- [55] 广州市统计局. 广州统计年鉴 2013. 北京: 中国统计出版社, 2013.
- [56] Chan M, Johansson MA. The incubation periods of dengue viruses. *PLOS ONE*, 2012, 7(11): e50972. DOI: 10.1371/journal.pone.0050972.
- [57] Brady OJ, et al. Modelling adult *Aedes aegypti* and *Aedes albopictus* survival at different temperatures. *Parasit Vectors*, 2013, 6: 351. DOI: 10.1186/1756-3305-6-351.
- [58] Kingma DP, Ba J. Adam: a method for stochastic optimization. In: *ICLR*, 2015. arXiv: 1412.6980.
- [59] Cranmer M, et al. Discovering symbolic models from deep learning with inductive biases. *NeurIPS*, 2023, 36: 17429–17442. DOI: 10.48550/arXiv.2006.11287.
- [60] Kraemer MUG, et al. Past and future spread of the arbovirus vectors *Aedes aegypti* and *Aedes albopictus*. *Nat Microbiol*, 2019, 4(5): 854–863. DOI: 10.1038/s41564-019-0376-y.
- [61] Chen RTQ, et al. Neural ordinary differential equations. In: *NeurIPS*, 2018, 31: 6571–6583. arXiv: 1806.07366.
- [62] Lowe R, et al. Nonlinear and delayed impacts of climate on dengue risk in Barbados. *PLOS Medicine*, 2018, 15(7): e1002613. DOI: 10.1371/journal.pmed.1002613.
- [63] Roberts DR, et al. Cross-validation strategies for data with temporal, spatial, hierarchical structure. *Ecography*, 2017, 40(8): 913–929. DOI: 10.1111/ecog.02881.
- [64] ANDERSON R M, MAY R M. *Infectious Diseases of Humans: Dynamics and Control*[M]. Oxford: Oxford University Press, 1991.
- [65] KEELING M J, ROHANI P. *Modeling Infectious Diseases in Humans and Animals*[M]. Princeton: Princeton University Press, 2008.
- [66] FOCKS D A, BRENNER R J, HAYES J, 等. Transmission thresholds for dengue in terms of *Aedes aegypti* pupae per person with discussion of their utility in source reduction efforts[J]. *American Journal of Tropical Medicine and Hygiene*, 2003, 68(6): 682–692.
- [67] VAN DEN DRIESSCHE P, WATMOUGH J. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission[J]. *Mathematical Biosciences*, 2002, 180(1–2): 29–48. DOI:10.1016/S0025-5564(02)00108-6.



- [68] BUTCHER J C. Numerical Methods for Ordinary Differential Equations[M]. 3rd ed. Chichester: John Wiley & Sons, 2016.
- [69] HUBER P J. Robust estimation of a location parameter[J]. Annals of Mathematical Statistics, 1964, 35(1): 73–101.
- [70] RAISSI M, PERDIKARIS P, KARNIADAKIS G E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations[J]. Journal of Computational Physics, 2019, 378: 686–707. DOI:10.1016/j.jcp.2018.10.045.
- [71] BRENT R P. Algorithms for Minimization without Derivatives[M]. Englewood Cliffs, NJ: Prentice-Hall, 1973.
- [72] LAI S, HUANG Z, ZHOU H, 等. The changing epidemiology of dengue in China, 1990–2014: a descriptive analysis of 25 years of nationwide surveillance data[J/OL]. BMC Medicine, 2015, 13: 100. DOI:10.1186/s12916-015-0336-1.
- [73] BURNHAM K P, ANDERSON D R. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach[M]. 2nd ed. New York: Springer, 2002.
- [74] HASTIE T, TIBSHIRANI R, FRIEDMAN J. The Elements of Statistical Learning[M]. 2nd ed. New York: Springer, 2009.
- [75] STORN R, PRICE K. Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces[J]. Journal of Global Optimization, 1997, 11(4): 341–359. DOI:10.1023/A:1008202821328.
- [76] BRADY O J, GOLDING N, PIGOTT D M, 等. Global temperature constraints on *Aedes aegypti* and *Ae. albopictus* persistence and competence for dengue virus transmission[J/OL]. Parasites & Vectors, 2014, 7: 338. DOI:10.1186/1756-3305-7-338.
- [77] LIU K, SUN J, LIU X, 等. Spatiotemporal patterns and determinants of dengue at county level in China from 2005–2017[J/OL]. International Journal of Infectious Diseases, 2018, 77: 96–104. DOI:10.1016/j.ijid.2018.09.028.

附录一 论文涉及的图表补充

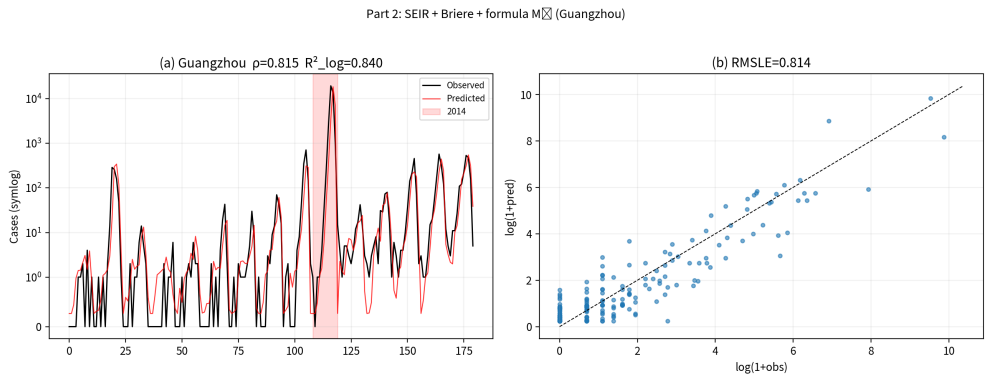


图 11: 广州市月度病例拟合曲线 (SEIR + Brière + 公式  $\hat{M}$  模型)

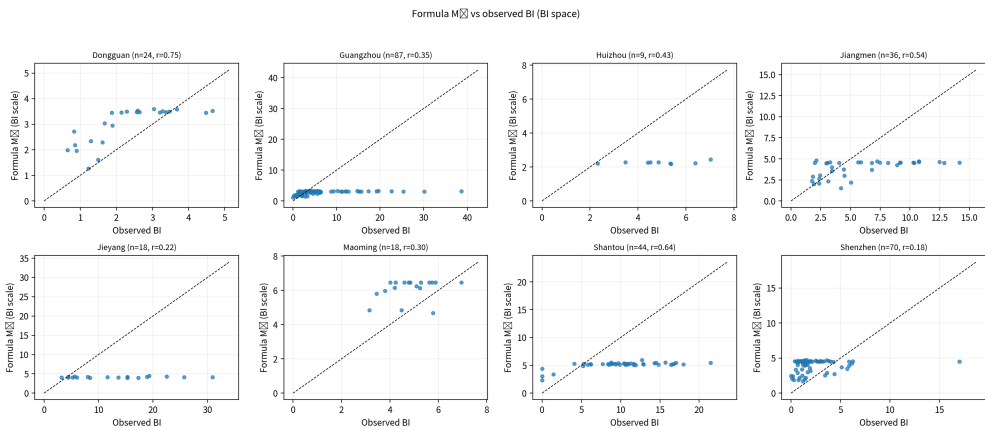


图 12: 公式  $\hat{M}$  与实测 BI 对比 (8 个有 BI 数据的城市)

## 致 谢

时光荏苒，研究生阶段的学习即将画上句号。回顾这段充实而难忘的岁月，心中充满感激。

首先，我要衷心感谢我的导师。在整个研究过程中，导师给予了我悉心的指导和无私的帮助。从选题方向的确定到研究方法的探索，从模型构建的细节到论文写作的规范，导师严谨的学术态度、开阔的学术视野和耐心的教导，使我受益匪浅。导师不仅在学术研究方面给予了我系统的训练，在跨学科思维和科学方法论方面也对我产生了深远的影响。

其次，感谢实验室的各位同学和师兄师姐。在数据收集、模型调试和结果讨论的过程中，大家给予了我许多建设性的意见和热情的帮助。特别感谢在符号回归算法调试和多城市数据预处理过程中提供技术支持的同学们，你们的协助使得本研究得以顺利推进。

感谢家人一直以来的理解和支持。在漫长的研究过程中，是你们的关爱和鼓励让我能够全身心投入学术研究。每一次遇到困难想要放弃时，是家人的陪伴给予了我坚持下去的力量。

感谢中国疾病预防控制中心和广东省疾控中心提供的病例报告数据，感谢美国国家海洋和大气管理局（NOAA）提供的开放气象数据资源。开放数据共享精神是推动科学进步的重要力量。

最后，感谢论文评审专家在百忙之中审阅本文并提出宝贵意见。你们的专业建议使论文质量得到了显著提升。

谨以此文献给所有关心和帮助过我的人。