·流行病与统计方法·

基于 Python 的不同时间序列模型在上海市宝山区其他感染性腹泻病预测中的比较

宗莲，郑家乐，向伦辉，何凡，刘效峰，朱江
上海市宝山区疾病预防控制中心（上海市宝山区卫生健康监督所），上海 201901

摘要：目的 通过比较季节性自回归滑动平均（seasonal autoregressive integrated moving awerage,SARIMA）模型、Prophet 模型、随机森林模型、极端梯度提升（eXtreme gradient boosting,XGBoost）模型和支持向量回归（support vector regression, SVR）模型对上海市宝山区其他感染性腹泻病发病率的预测效果，探索最优预测模型。**方法** 利用 2014—2022 年上海市宝山区其他感染性腹泻病月发病率作为训练集，运用 Python 语言分别构建 SARIMA 模型、Prophet 模型、含与不含气象特征的随机森林模型、含与不含气象特征的 XGBoost 模型以及含与不含气象特征的 SVR 模型，对 2023—2024 年月发病率进行预测。通过比较预测值与实际值的平均绝对误差（mean absolute error,MAE）、均方根误差（root mean square error,RMSE）和平均绝对百分比误差（mean absolute percentage error,MAPE），遴选出最优模型。**结果** 2014—2024 年上海市宝山区共报告其他感染性腹泻病 6 868 例，年均发病率 28.81/10 万，病例呈双峰季节性分布（夏季和冬春季高发）。通过比较不同时间序列模型对 2023—2024 年月发病率的预测效果，不含气象特征的 XGBoost 模型拟合精度指标综合表现最优（MAE=5.089, RMSE=6.844, MAPE=54.646%）。**结论** 不含气象特征的 XGBoost 模型为宝山区其他感染性腹泻病短期预测的最佳模型，可为宝山区其他感染性腹泻病的预警和防控提供科学依据。

关键词：Python；时间序列模型；其他感染性腹泻病；预测
中图分类号：R183.4 **文献标志码**：A **文章编号**：1003-8507（2026）07-1172-09
DOI：10.20043/j.cnki.MPM.202507165

Comparison of different Python-based time series models in the prediction of other infectious diarrhea diseases in Baoshan District, Shanghai

ZONG Lian, ZHENG Jia-le, XIANG Lun-hui, HE Fan, LIU Xiao-feng, ZHU Jiang
Shanghai Baoshan District Center for Disease Control and Prevention (Shanghai Baoshan District Health Supervision Institute), Shanghai 201901, China

Abstract: Objective To compare the predictive performance of seasonal autoregressive integrated moving average (SARIMA) model, Prophet model, random forest (RF) model, extreme gradient boosting (XGBoost) model, and support vector regression (SVR) model on the incidence of other infectious diarrhea diseases in Baoshan District, Shanghai, and to explore the optimal forecasting model. **Methods** Using the monthly incidence rates of other infectious diarrhea diseases in Baoshan District, Shanghai from 2014 to 2022 as the training set, the SARIMA model, the Prophet model, the RF model with and without meteorological features, the XGBoost model with and without meteorological features, and the SVR model with and without meteorological features were constructed in Python to predict the monthly incidence rates from 2023 to 2024. The optimal model was selected by comparing the mean absolute error (MAE), root mean square error (RMSE) and mean absolute percentage error (MAPE) of the predicted and actual values. **Results** From 2014 to 2024, a total of 6 868 cases of other infectious diarrheal diseases were reported in Baoshan District, Shanghai, with an average annual incidence rate of 28.81 per 100 000 population. The cases exhibited a bimodal seasonal distribution (peaking in summer and winter-spring). By comparing the predictive performance of different time series models on the monthly incidence rate from 2023 to 2024, the XGBoost model without meteorological features showed the best comprehensive performance in terms of fitting accuracy indicators (MAE=5.089, RMSE=6.844, MAPE=54.646%). **Conclusion** The XGBoost model without meteorological features is the best model for short-term prediction of other infectious diarrhea diseases in Baoshan District, which can provide a scientific basis for early warning and prevention of other infectious diarrhea diseases in Baoshan District.

基金项目：上海市加强公共卫生体系建设三年行动计划（2023—2025 年）重点学科（GWVI-11.1-06）；上海市宝山区科学技术委员会医学卫生项目（2025-E-11）；宝山区传染病流行病学重点学科项目（BSZK-2023-A20）；宝山区卫生健康委优青（育才）计划（BSWSYC-2023-21）
作者简介：宗莲（1990—），女，硕士，主管医师，研究方向：疾病预防控制
通信作者：朱江，E-mail：zhujiang@bscdc.org.cn

Keywords: Python; Time series model; Other infectious diarrhea diseases; Forecast

其他感染性腹泻病是由多种病原体引起的，除霍乱、细菌性或阿米巴性痢疾、伤寒及副伤寒外以腹泻为主要症状的法定丙类传染病[1]。监测数据显示，2024 年上海市其他感染性腹泻病已超越手足口病跃居丙类传染病发病数第二位。因此，寻求有效的时间序列模型对其他感染性腹泻病进行早期预警具有重要的现实意义。时间序列模型作为分析和预测时序数据的一类统计或机器学习方法[2]，已被广泛应用于其他感染性腹泻病发病的预测研究[3-4]，但针对不同模型预测效果比较的系统性研究较为有限。本研究利用上海市宝山区 2014 年 1 月—2024 年 12 月其他感染性腹泻病的报告信息分别建立季节性自回归滑动平均（seasonal autoregressive integrated moving average，SARIMA）模型、Prophet 模型、随机森林模型、极端梯度提升（eXtreme gradient boosting，XGBoost）模型和支持向量回归（support vector regression，SVR）模型，通过评估在测试集上的预测效果，遴选出最适合宝山区其他感染性腹泻病发病趋势预测的模型，从而为本地疫情的早期预警、动态监测和防控处置提供科学依据。

1　资料与方法

1.1　资料来源　其他感染性腹泻病发病资料来自中国疾病预防控制信息系统，为 2014 年 1 月—2024 年 12 月报告、现住址为上海市宝山区的所有已审核个案。气温和降水量数据来源于美国国家海洋和大气管理局下设的国家环境信息中心；人口资料来源于上海市宝山区统计年鉴。

1.2　研究方法

1.2.1　SARIMA 模型　SARIMA 模型在自回归滑动平均（autoregressive integrated moving average，ARIMA）模型基础上，增加了季节性参数，专门用于分析预测具有季节性等周期规律的时间序列数据，是一种在气象、经济、医疗卫生领域广泛应用的经典时间序列模型。该模型一般为 ARIMA$(p,d,q)(P,D,Q)s$ 的形式，p 为自回归阶数，d 为非季节性差分阶数，q 为滑动平均阶数；P 为季节性自回归阶数，D 为季节性差分阶数，Q 为季节性滑动平均阶数，s 为周期长度。本研究使用的为其他感染性腹泻病月发病数据，因此 s=12。建模步骤一般如下：（1）平稳化序列：观察原始时间序列的趋势和周期变化，对时间序列进行单位根（augmented Dickey-Fuller，ADF）检验，判断序列的平稳性，对不平稳序列采取差分使其平稳化；（2）模型参数选取：对参数 p、q、P、Q 建立参数网格进行搜索，依据赤池信息准则（Akaike information criterion，AIC）、贝叶斯信息准则（Bayesian information criterion，BIC）等，筛选出模型最优参数；（3）效果评估：对选出模型的残差进行 Ljung-Box 白噪声检验和正态性检验。

1.2.2　Prophet 模型　该模型是由 Facebook 开发的一个开源时间序列预测模型，于 2017 年发布，适合用于处理具有趋势性、季节性和节假日效应的数据[5]，其核心采用加法模型分解时间序列为趋势项、季节项和节假日项，表达式可写为：

$$y(t)=g(t)+s(t)+h(t)+\varepsilon(t)$$

其中 $g(t)$ 为趋势项，用于拟合非周期的趋势变化；$s(t)$ 为周期项，用傅里叶级数建立周期模型，以拟合季、月、周、日等周期性变化规律；$h(t)$ 为节假日项，用于处理节假日因素对预测的影响；$\varepsilon(t)$ 为噪声项，表示未预测到的随机扰动。本研究通过网格搜索方式调整参数组合，得到最佳的模型预测效果。

1.2.3　随机森林模型　随机森林模型为一种基于 Bagging 集成学习方法的机器学习模型，其本质是一个包含多个决策树的分类器。决策树是一个基本的分类或回归模型，可以通过系列特征将数据分割成不同类别或预测一个连续的数值[6]。随机森林模型通过构建多个决策树提高准确性和鲁棒性，可不受多重共线性特征影响，对非线性关系数据处理结果较好。模型算法主要是对每个决策树，进行有放回的随机抽样获得样本集，从所有的 M 个特征中随机选取 m 个特征（m≤M，默认 m=$\sqrt{M}$），在树的每个分裂点，根据所选特征进行最佳分裂，直到得到最终结果。在树的构建过程中，每棵树都会完整生长而不被剪枝。最后根据所有决策树的预测结果，进行投票或回归得到最终结果。本研究通过网格搜索方式遍历超参数组合，对每种组合进行训练和评估，选取性能指标最优的参数组合构建模型。

1.2.4　XGBoost 模型　XGBoost 模型基于梯度提升框架进行集成学习。它通过构建一系列决策树等弱预测模型，然后加权平均这些模型的预测结果来学习复杂的非线性关系，提高整体的预测性能[7]。模型将损失函数进行二阶泰勒展开，通过 L1 和 L2 正则化优化模型性能，防止过拟合。XGBoost 模型在每棵树的构建过程中使用了更精细的剪枝策略，在分裂点可并行评估所有分割方式。其目标函数表示为：

$$Obj=\sum_{i=1}^{n}L(y_i,\hat{y}_i)+\sum_{t=1}^{T}\Omega(f_t)$$

其中第一部分 $L(y_i,\hat{y}_i)$ 表示误差损失，用以衡量实际值和预测值的差异；第二部分 $\Omega(f_t)$ 表示正则化

项，通过叶子节点数和叶子节点权重等限制模型的复杂度。

1.2.5　SVR 模型　SVR 模型是机器学习领域用于回归问题的一种监督学习模型，相较于传统线性回归，SVR 模型可通过核方法捕捉非线性关系，常用的核函数包括线性核、高斯核、多项式核等。模型的核心目标是试图找到一个最优的超平面，使所有输入的数据点到这个平面的距离尽可能小，同时也引入正则化项控制模型的复杂度避免过拟合。其原理是在线性函数两侧设置一个间距为 ε 的间隔带，对所有落入间隔带内的样本不计算损失，最后通过最小化损失和最大化间隔来优化模型[8]。其目标函数为：

$$\min_{w,b,\xi_i,\hat{\xi}_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i, \hat{\xi}_i)$$

$$s.t.\quad f(x_i) - y_i \le \varepsilon + \xi_i$$

$$y_i - f(x_i) \le \varepsilon + \hat{\xi}_i$$

$$\xi_i,\ \hat{\xi}_i \ge 0,\quad i = 1,\ 2,\ \ldots\ldots,\ m$$

其中 w、b 为超平面参数，C 为正则化参数，控制模型的复杂度和误差；ξ_i、$\hat{\xi}_i$ 为松弛变量，表示允许超出 ε 带的范围。本研究通过网格搜索和交叉验证方法来寻求最佳参数来构建最适合预测发病率的模型。

1.2.6　模型评价和选取　本研究利用 2014 年 1 月—2022 年 12 月上海市宝山区其他感染性腹泻病月发病率作为训练集，并用 2023 年 1 月—2024 年 12 月的月发病率作为测试集，通过比较不同时间序列模型的预测值和实际值的平均绝对误差（mean absolute error，MAE）、均方根误差（root mean square error，RMSE）、平均绝对百分比误差（mean absolute percentage error，MAPE），最后遴选出最适合预测上海市宝山区其他感染性腹泻病发病率趋势的模型。

1.3　统计学分析　应用 Microsoft Excel 2019 进行数据处理和描述性统计，分别使用 Python 3.9 中的 statsmodels 库进行数据平稳性检验、SARIMA 模型构建和互相关函数（cross-correlation function，CCF）分析，scikit-learn 库进行网格搜索、随机森林模型和 SVR 模型的构建以及精度指标的计算，prophet 库进行 Prophet 模型构建、xgboost 库进行 XGBoost 模型的构建、matplotlib 库进行数据可视化。检验水准 α=0.05。

2　结　果

2.1　原始序列分析　2014—2024 年，宝山区共报告其他感染性腹泻病 6 868 例，年发病率在 10.21/10 万（2022 年）~ 48.56/10 万（2019 年）间波动，年均报告发病率为 28.81/10 万。月份分布上看，整体呈现双峰分布，两个高峰分别出现在 6—9 月（37.57%）及 12 月—次年 3 月（35.06%）。分别绘制 2014 年 1 月—2024 年 12 月宝山区月平均气温、降水量与其他感染性腹泻病发病率关系的散点图，见图 1。Spearman 相关性分析显示，月平均气温与发病率的相关系数 ρ=0.109（95%*CI*: −0.072 ~ 0.289，*P*=0.212），降水量与发病率的相关系数 ρ=0.075（95%*CI*: −0.105 ~ 0.248，*P*=0.391），均呈微弱正相关，但无显著统计学意义。

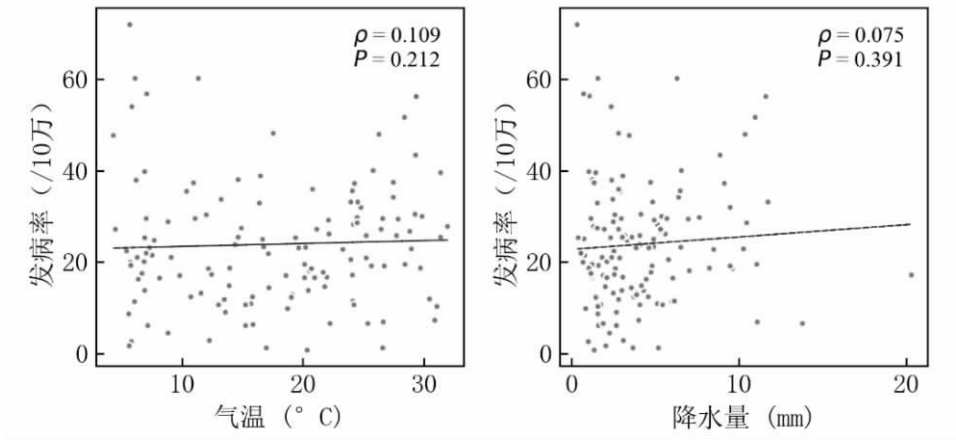


图 1　气温、降水量与其他感染性腹泻病发病率的关系

Figure 1　Relationship between temperature, precipitation and incidence rate of other infectious diarrhea diseases

2.2　时间序列模型构建

2.2.1　SARIMA 模型构建　利用 seasonal_decompose 函数对 2014—2024 年上海市宝山区其他感染性腹泻病月度发病率时间序列进行季节性分解，绘制分解序列图，观察其总体趋势、季节波动和残差分布情况，2020 年之前宝山区其他感染性腹泻病发病率总体较为平稳，2020 和 2022 年有明显的降低。月度发病率存在明显季节性，发病率时间序列的残差在一定范围内波动。见图 2。

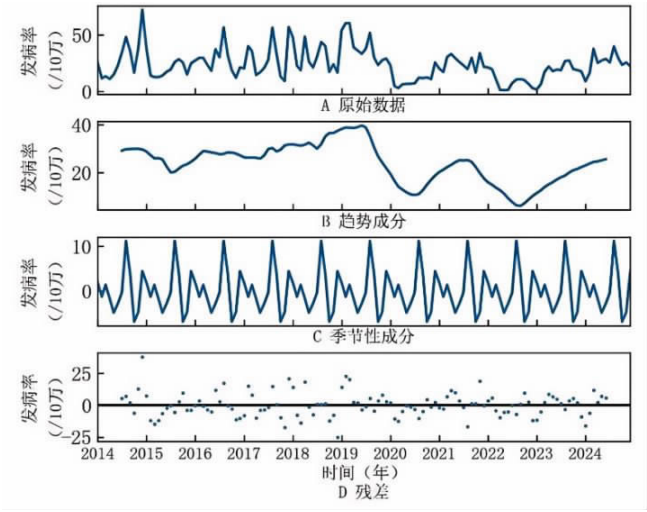


图 2　2014—2024 年上海市宝山区其他感染性腹泻病例时间序列分解图

Figure 2　Time series decomposition of other infectious diarrhea diseases cases in Baoshan District, Shanghai from 2014 to 2024

划分 2014 年 1 月—2022 年 12 月数据为训练集，2023 年 1 月—2024 年 12 月数据为测试集。利用 adfuller 函数对发病率时间序列进行平稳性检验，Dickey–Fuller 统计量为 −3.232（P=0.018），为平稳序列，无需差分。考虑到数据有明显的季节性周期，对序列进行一阶季节性差分，因此选取模型参数 d=0，D=1。为寻找最优的 p、q、P、Q 参数，利用网格搜索方法并结合 SARIMAX 函数构建模型，根据 AIC 准则，最终选取 p=1，q=4，P=2，Q=0。ARIMA（1, 0, 4）(2, 1, 0)$_{12}$ 模型 AIC 值为 564.936。对模型残差绘制密度直方图、正态 Q–Q 图和自相关图进行可视化诊断，同时进行 Ljung–Box 白噪声检验（χ^2=0.211，P=0.646），差异无统计学意义，残差为白噪声序列。见图 3。

2.2.2　Prophet 模型构建　对 Prophet 模型的参数建立网格，由于发病率为月度数据，且有明显季节性，因此 yearly_seasonality 参数选择为 True。利用训练集数据通过网格搜索用 Prophet 函数构建模型，进行交叉验证，并根据精度指标 MAE 最小的原则选取最优参数，见表 1。用最佳参数重新训练模型，用 add_country_holidays 方法加入中国法定节假日作为节假日参数，对 2023—2024 年月发病率进行预测和效果评价。对最优模型的残差进行 Ljung–Box 检验（χ^2=31.491，P<0.001），残差仍存在显著自相关。

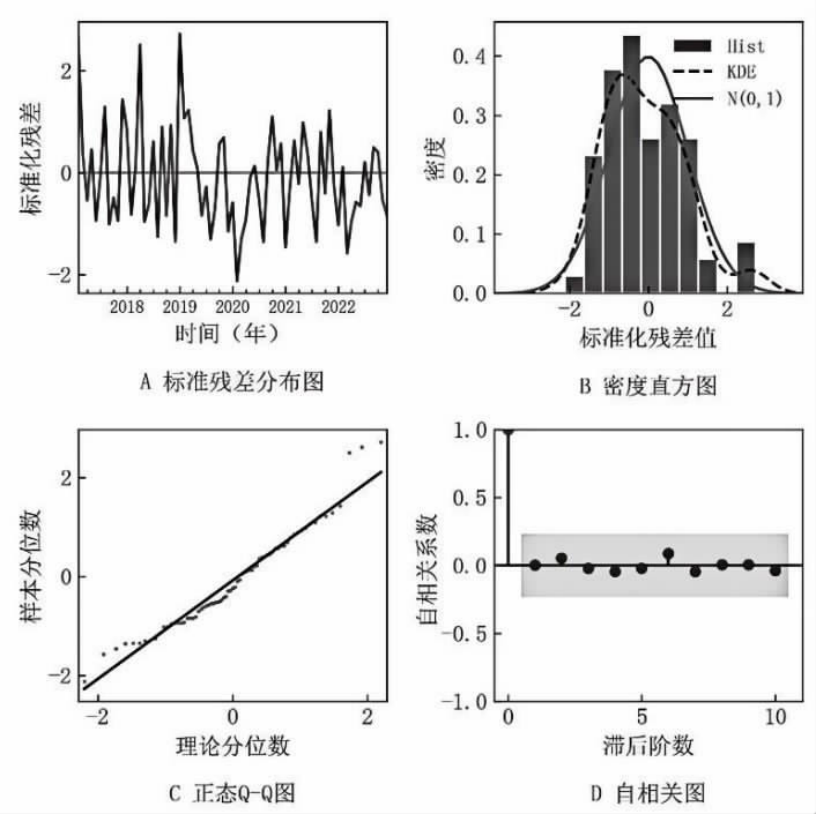


图 3　ARIMA(1, 0, 4)(2, 1, 0)$_{12}$ 模型残差诊断图

Figure 3　Diagnosis diagrams of the residuals in ARIMA(1,0,4)(2,1,0)$_{12}$ model

表 1　Prophet 模型参数网格和最优参数

Table 1　Parameters grid and optimal parameters of Prophet model

参数	参数范围	最优参数
changepoint_prior_scale	[0.1, 0.05, 0.1]	0.1
seasonality_prior_scale	[1.0, 5.0, 10.0]	1.0
holidays_prior_scale	[10.0, 15.0]	10.0
seasonality_mode	[‘additive’, ‘multiplicative’]	‘multiplicative’

2.2.3　随机森林模型构建　通过 CCF 分别检测上海市宝山区的月平均气温、降水量与其他感染性腹泻病月发病率之间的滞后相关性，显示月均气温的 1～3 阶滞后与发病率均未呈现统计学显著相关性（$P>0.05$），降水量的 1 阶滞后与发病率之间的相关系数达到显著性水平（$P=0.046$）。分别构建含气温和降水特征、仅含气温特征、仅含降水特征和不含气象特征的随机森林模型，对模型参数建立网格，使用 Random Forest Regressor 方法初始化模型，设置交叉验证参数 cv=5，使用 GridSearchCV 函数分别进行网格搜索并选取最优参数，见表 2。通过比较各个模型的 MAE，不含气象特征的随机森林模型预测效果更佳，见图 4。对最佳随机森林模型的残差进行 Ljung-Box 检验（$\chi^2=0.356$，$P=0.551$），差异无统计学意义，残差为白噪声序列。

表 2　随机森林模型参数网格和最优参数

Table 2　Parameters grid and optimal parameters of random forest model

参数	参数范围	最优参数			
		含气温和降水特征	仅含气温特征	仅含降水特征	不含气象特征
n_estimators	[100, 200, 300]	100	100	100	100
max_depth	[None, 10, 20, 30]	10	None	None	10
min_samples_leaf	[2, 5, 10]	2	2	2	2
min_samples_split	[2, 5, 10]	5	5	5	2
max_features	['sqrt', 'log2', None]	None	None	None	None

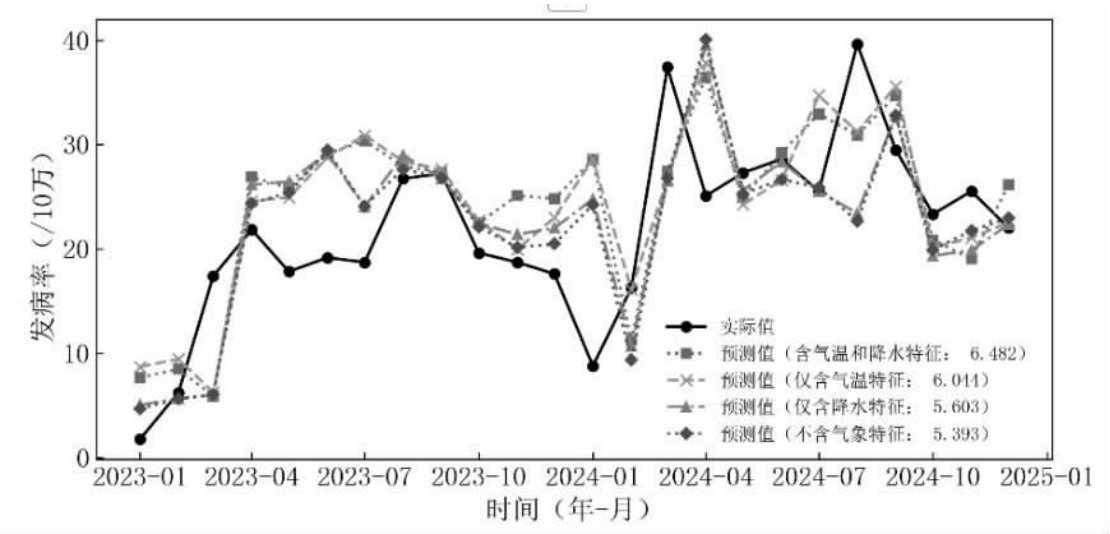


图 4　含不同特征工程的最优随机森林模型比较

Figure 4　Comparison of optimal random forest models containing different feature engineering

2.2.4　XGBoost 模型构建　对上海市宝山区其他感染性腹泻病的月发病率序列分别构建含气温和降水特征、仅含气温特征、仅含降水特征和不含气象特征的 XGBoost 模型，对模型的参数建立网格，使用 XGBRegressor 方法初始化模型，设置交叉验证参数 cv=5，用 GridSearchCV 函数进行网格搜索并选取最优参数，见表 3。通过比较各个模型的 MAE，不含气象特征的 XGBoost 模型预测效果更佳，见图 5。对最佳 XGBoost 模型的残差进行 Ljung-Box 检验（$\chi^2=12.878$，$P<0.001$），残差仍存在显著自相关。

表 3　XGBoost 模型参数网格和最优参数

Table 3　Parameters grid and optimal parameters of XGBoost model

参数	参数范围	最优参数			
		含气温和降水特征	仅含气温特征	仅含降水特征	不含气象特征
max_depth	[3, 5, 7]	3	3	7	3
learning_rate	[0.01, 0.1, 0.2]	0.01	0.01	0.01	0.01
n_estimators	[100, 200, 300]	100	100	100	100
subsample	[0.7, 0.9]	0.7	0.7	0.7	0.7
colsample_bytree	[0.7, 0.9]	0.7	0.9	0.7	0.9
gamma	[0, 0.1]	0	0	0.1	0
min_child_weight	[1, 3]	3	3	3	3

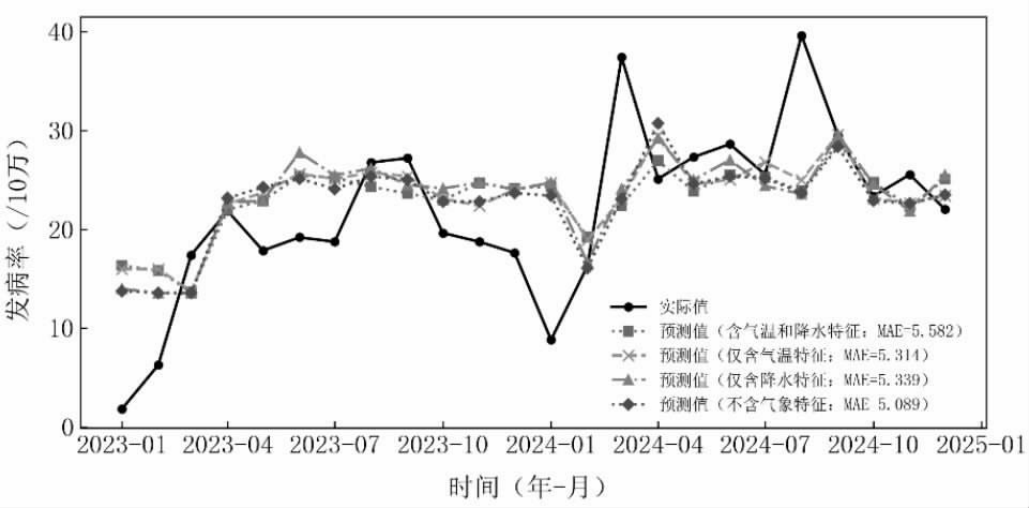


图 5　含不同特征工程的最优 XGBoost 模型比较

Figure 5　Comparison of optimal XGBoost models containing different feature engineering

2.2.5　SVR 模型构建　对上海市宝山区其他感染性腹泻病的月发病率序列分别构建含气温和降水特征、仅含气温特征、仅含降水特征和不含气象特征的 SVR 模型，使用 Standard Scaler 方法对特征进行标准化。对 SVR 模型的参数建立网格，使用 SVR 函数初始化模型，设置交叉验证参数 cv=5，用 GridSearchCV 函数进行网格搜索并选取最优参数，见表 4。通过比较各个模型的 MAE，不含气象特征的 SVR 模型预测效果更佳，见图 6。对最佳 SVR 模型的残差进行 Ljung-Box 检验（χ^2=34.017，P<0.001），残差仍存在显著自相关。

表 4　SVR 模型参数网格和最优参数

Table 4　Parameters grid and optimal parameters of SVR model

参数	参数范围	最优参数			
		含气温和降水特征	仅含气温特征	仅含降水特征	不含气象特征
C	[0.1, 1, 10, 100]	10	10	10	10
gamma	[0.01, 0.1, 1, ‘scale’, ‘auto’]	‘auto’	‘auto’	0.1	0.1
epsilon	[0.01, 0.1, 0.5, 1]	1	1	1	1
kernel	[‘rbf’, ‘linear’, ‘poly’]	‘rbf’	‘rbf’	‘rbf’	‘rbf’

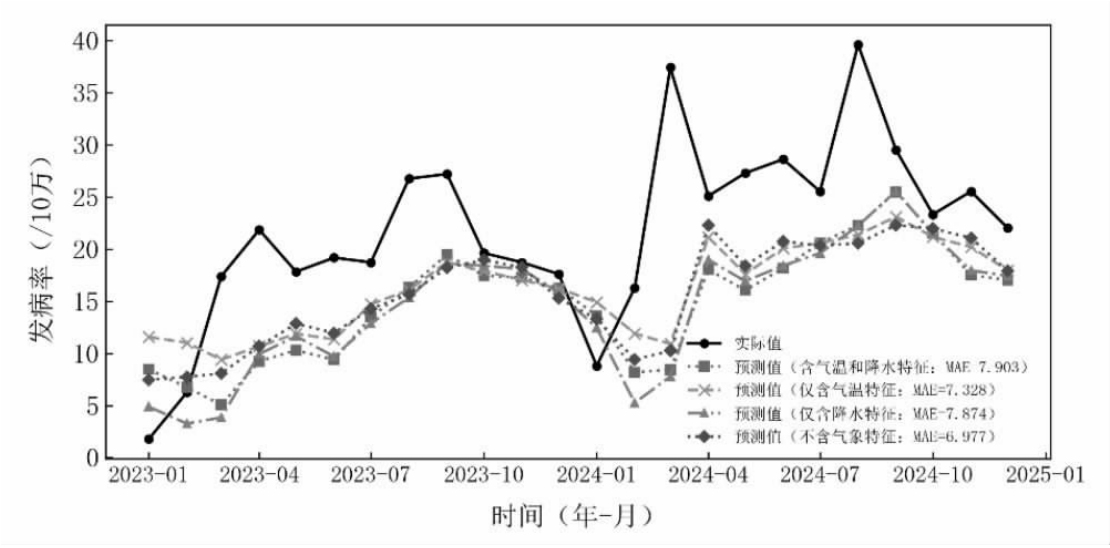
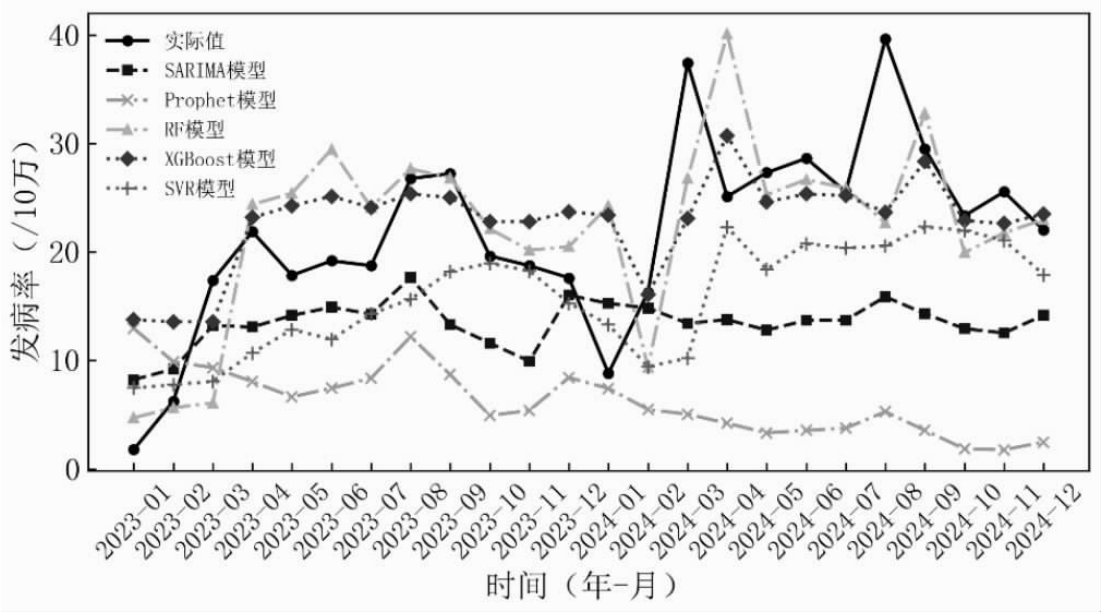


图 6　含不同特征工程的最优 SVR 模型比较

Figure 6　Comparison of optimal SVR models containing different feature engineering

2.2.6　最佳预测模型选取　分别选取上述不同类型　　的最优时间序列模型预测 2023 年 1 月—2024 年 12

月上海市宝山区其他感染性腹泻病的月发病率情况，并与实际值进行比较，通过计算 MAE、RMSE 和 MAPE，综合比较可见不含气象特征的 XGBoost 模型预测效果较其他模型更为优秀（MAE=5.089，RMSE=6.844，MAPE=54.646%）。见表 5 和图 7。



注：RF 模型为随机森林模型。

图 7　不同时间序列模型预测效果对比

Figure 7　Comparison of prediction performance of different time series models

表 5　不同时间序列模型精度系数比较

Table 5　Comparison of accuracy coefficients of different time series models

模型	MAE	RMSE	MAPE(%)
SARIMA 模型	11.023	12.637	71.253
Prophet 模型	16.736	18.643	93.961
随机森林模型	5.393	7.390	34.417
XGBoost 模型	5.089	6.844	54.646
SVR 模型	6.977	9.108	41.782

3　讨　论

研究显示，随机森林、XGBoost 和 SVR 模型的预测性能总体优于 SARIMA 模型和 Prophet 模型，这可能与机器学习模型在处理非线性关系和特征交互上的固有优势有关，这也与诸多研究的结果一致[9-10]。SARIMA 模型虽然能捕捉时间序列的趋势和季节性，但是对复杂的非线性关系建模能力有限，Prophet 模型能灵活处理趋势变化、周期规律和节假日效应，但在处理非线性效应和外部变量关系时稍逊一筹，导致了在测试集上表现欠拟合[11]。相比之下，随机森林模型和 XGBoost 模型可通过决策树的递归分割，对变量间的非线性交互效应自动建模；SVR 模型则通过核函数将数据映射到高维空间，拟合复杂的非线性边界[12]。本研究中，XGBoost 模型较随机森林模型和 SVR 模型表现出更低的 MAE 与 RMSE 值，预测精度更为优秀，这可能与模型的固有机制相关。XGBoost 模型采用梯度提升框架，通过序列化地构建决策树并专注于修正前序模型的残差，使其更擅长捕捉时间序列中复杂的非线性模式，其内置的正则化项有效控制了模型复杂度，避免了过拟合，从而获得了更好的泛化性能。而随机森林模型的 Bagging 策略虽能有效降低方差，但其精度上限可能不如 Boosting 策略；而 SVR 的性能则更依赖于核函数与超参数的精细调优[13]。

SARIMA 模型作为传统时间序列模型，进行残差诊断是模型构建的必要步骤，其目标是验证模型是否完全提取了数据中的线性依赖关系；而机器学习模型更关注预测精度，将残差分析视为一种辅助诊断工具而非强制性步骤，其通过正则化、交叉验证、集成学习等不同机制来控制过拟合，而非完全依赖残差的独立性假设。为了体现研究的严谨性，对所有采用模型的残差进行了 Ljung-Box 检验，显示 SARIMA、随机森林模型的残差序列为白噪声，而 Prophet、XGBoost 和 SVR 模型的残差序列均还存在自相关。但基于机器学习模型显著更低的 MAE 值和 RMSE 值，它们仍然是更优秀的预测模型，残差中剩余的自相关信息对预测精度的提升潜力已非常有限。

本研究在构建不同类型的时间序列模型时，用网格搜索的方法选取超参数，存在着一定的局限性[14]。一是，网格搜索本质是在预设的离散参数组合下进行

穷举搜索，参数组合的数量会随着参数数量和预设值数量呈指数级增长，计算成本会因为参数组合爆炸和时序重新训练验证急剧升高，因此需在搜索效率和模型性能之间做出平衡；二是，参数网格为离散化采样，当最优参数位于网格点之间时，易遗漏连续参数的最优解，且无法捕捉到有些关键参数之间的协同和交互作用；另外网格搜索对参数的静态优化，难以捕捉其他感染性腹泻病在真实世界中存在的局部暴发和波动特征，虽然在测试集上表现较好，但对长期预测的泛化能力仍存在不足。

尽管 Spearman 相关性分析显示月平均气温、降水量与其他感染性腹泻发病率之间均无显著的统计学关联，但也未排除可能存在复杂的非线性关联。基于 CCF 分析，本研究在特征工程中引入了平均气温、降水量及降水量的 1 阶滞后项，通过构建不同特征组合的模型，评估并对比其精度指标，以确定最优特征集。虽然大量文献证实了气温与降水对感染性腹泻发病存在相关性[15-16]，但本研究发现在引入这些气象特征后，所有机器学习模型的预测性能均未得到改善，反而有所下降。这一的结果可能由以下原因导致：首先，时间特征是气象季节性的强代理变量，模型已通过月份特征间接学习到了气象的季节性变化信号，引入的气温和降水特征反而成为了冗余信息；其次，气象因素的影响可能存在滞后效应与非线性的复杂机制，简单的月均气温可能掩盖了短期的气温波动对其他感染性腹泻病发病率的影响，而未能有效刻画这一关系；最后，其他感染性腹泻病的发病可能还会受到更多因素的影响，如病原谱变化、新冠疫情、公众健康行为改变等因素，可能掩盖了气象因素的独立贡献。这一发现提示，在利用机器学习模型进行流行病学预测时，应审慎使用依赖于固定特征的预测模型。未来的预警系统应着眼于动态和自适应型，能够定期评估各预测因子的效用，并在信号减弱时自动降权或剔除该因子。在今后的研究中，也可尝试构建混合模型[17]，采用随机搜索、贝叶斯优化等方法寻找最佳参数[18]，并引入更多的环境因子和社会特征[19]，来提高模型的精准性和鲁棒性，为宝山区其他感染性腹泻病提供更可靠有效的预警参考。

利益冲突声明　本研究不存在任何利益冲突

参考文献

[1] 翟雪艳，高洪朱，莫颖宁. 2010-2020 年中国其他感染性腹泻病的流行病学分析 [J]. 中华疾病控制杂志，2025，29（1）：34-41.
Zhai XY, Gao HZ, Mo YN. Epidemiological analysis of other infectious diarrhea diseases in China from 2010 to 2020[J]. Chinese Journal of Disease Control & Prevention, 2025, 29 (1): 34-41.(In Chinese)

[2] 李长生. SARIMA 与 Prophet 的混合算法在时间序列预测中的应用研究[J]. 软件，2025，46（1）：7-9.
Li CS. A hybrid model based on SARIMA and prophet for improving time series forecasting accuracy [J]. Computer Engineering & Software, 2025, 46(1): 7-9.(In Chinese)

[3] 茅蓉，向泽林，王远航，等. 2017-2022 年浙江省其他感染性腹泻病 ARIMA 模型预测精度分析 [J]. 健康研究，2024，44（5）：528-533.
Mao R, Xiang ZL, Wang YH, et al. Analysis of the prediction accuracy of ARIMA model for other infectious diarrhea diseases in Zhejiang Province from 2017 to 2022 [J]. Health Research, 2024, 44(5): 528-533.(In Chinese)

[4] 徐海涛，陈颖峰，唐智峰，等. 时间序列分析在其它感染性腹泻预测预警中的应用 [J]. 中国卫生检验杂志，2012，22（12）：2945-2947.
Xu HT, Chen YF, Tang ZF, et al. The application of time series analysis in forecast and warning of other infectious diarrhea [J]. Chinese Journal of Health Laboratory Technology, 2012, 22(12): 2945-2947.(In Chinese)

[5] 姜雨淇，龙江，赵金华，等. 2017-2023 年青海省其他感染性腹泻病发病率的时空分析及预测模型比较[J]. 中华疾病控制杂志，2024，28（11）：1301-1307.
Jiang YQ, Long J, Zhao JH, et al. The spatio-temporal analysis and prediction model comparison of incidence rate of other infectious diarrhea diseases in Qinghai Province from 2017 to 2023 [J]. Chinese Journal of Disease Control & Prevention, 2024, 28 (11): 1301-1307.(In Chinese)

[6] Fang X, Liu W, Ai J, et al. Forecasting incidence of infectious diarrhea using random forest in Jiangsu Province,China [J]. BMC Infectious Diseases, 2020, 20(1): 222.

[7] Liang D, Wang L, Liu S, et al. Global incidence of diarrheal diseases-an update using an interpretable predictive model based on XGBoost and SHAP:a systematic analysis [J]. Nutrients, 2024, 16 (18): 3217.

[8] 霍静，王永明，顾君忠. 感染性腹泻周发病例数的 PCA-SVM 回归预测研究[J]. 计算机应用与软件，2016（2）：51-54.
Huo J, Wang YM, Gu JZ. Research on PCA-SVM regressive prediction of weekly cases of infectious diarrhea[J]. Computer Applications and Software, 2016(2): 51-54.(In Chinese)

[9] 黄蓉蓉. 不规则采样季节性传染病时间序列预测方法比较[D]. 沈阳：中国医科大学，2024.
Huang RR. Comparison of forecasting methods for irregularly sampled seasonal infectious disease time series[D]. Shenyang: China Med University, 2024.(In Chinese)

[10] 郝强. 气象因素对其他感染性腹泻发病的影响及预测预警研究[D]. 济南：山东大学，2022.
Hao Q. Impact of meteorological factors on other infectious diarrhea and forecasting and early warning [D]. Jinan: Shandong Univ, 2022.(In Chinese)

[11] 祖力皮卡尔·吐迪，杨振，卢耀勤，等. SARIMA 等模型在乌鲁木齐市其他感染性腹泻病发病预测中的应用[J]. 医学动物防制，2024，40（5）：435-439.
Zu Li Pi Ka Er - Tu Di, Yang Z, Lu YQ, et al. Application of

（下转第 1194 页）

and bacterial antimicrobial resistance in China: a systematic analysis for the global burden of disease study 2019 [J]. Lancet Reg Health West Pac, 2024, 43: 100972.

[19] Kreisel KM, Llata E, Haderxhanaj L, et al. The burden of and trends in pelvic inflammatory disease in the United States, 2006–2016[J]. The Journal of Infectious Diseases, 2021, 224(12): S103–S112.

[20] Kenyon C, Herrmann B, Hughes G, et al. Management of asymptomatic sexually transmitted infections in Europe: towards a differentiated, evidence–based approach [J]. Lancet Reg Health Eur, 2023, 34: 100743.

[21] Soe NN, Yu Z, Latt PM, et al. Evaluation of artificial intelligence–powered screening for sexually transmitted infections–related skin lesions using clinical images and metadata[J]. BMC Medicine, 2024, 22(1): 296.

收稿日期:2025–09–08

（上接第 1179 页）

SARIMA and other models in the incidence prediction of other infectious diarrhea diseases in Urumqi [J]. Journal of Medical Pest Control,2024, 40(5): 435–439.(In Chinese)

[12] 周慧娴,龙前进,邓春燕,等. 传染病时间序列分析在传染病监测和早期预警中的应用[J]. 现代疾病预防控制,2025,36(2):159–164.
Zhou HX, Long QJ, Deng CY, et al. Application of time series analysis in infectious diseases surveillance and early warning system[J]. Modern Disease Control and Prevention, 2025, 36 (2): 159–164.(In Chinese)

[13] 李锡辉,时圆圆,祝光湖. 基于机器学习和时间序列方法的传染病规模预测模型[J]. 桂林电子科技大学学报,2024,44(1):87–92.
Li XH, Shi YY, Zhu GH. Prediction of notifiable infectious diseases based on the combination of machine learning and time series methods[J]. Journal of Guilin University of Electronic Technology, 2024, 44(1): 87–92.(In Chinese)

[14] 秦强. 基于组合模型的突发急性传染病疫情预测研究[D]. 武汉:武汉科技大学,2024.
Qin Q. Prediction of epidemic of emerging infectious diseases based on combination models [D]. Wuhan: Wuhan Technol Univ, 2024.(In Chinese)

[15] 汪君君,王太武,徐庆,等. 气温对南京市其他感染性腹泻病发病影响的时间序列研究 [J]. 中国卫生统计,2024,41(6):895–900.
Wang JJ, Wang TW, Xu Q, et al. A time–series study of the effect of temperature on the incidence of other infectious diarrhoeal diseases in Nanjing [J]. Chinese Journal of Health Statistics, 2024, 41 (6): 895–900.(In Chinese)

[16] 李燃, 陈煜. 2012—2021 年成都市气象因素对感染性腹泻发病的影响研究[J]. 成都医学院学报,2023,18(1):11–14.
Li R, Chen Y. Study on the influence of meteorological factors on the incidence of infectious diarrhea in Chengdu from 2012 to 2021[J]. Journal of Chengdu Medical College, 2023, 18 (1): 11–14.(In Chinese)

[17] 王立柱,吴品康,王秋萍. 随机森林模型参数寻优算法比较分析[J]. 沈阳师范大学学报:自然科学版,2024,42(5):420–426.
Wang LZ, Wu PK, Wang QP. Comparative analysis of parameter optimization algorithms in random forest models[J]. Journal of Shenyang Normal University(Natural Science Edition), 2024, 42(5): 420–426.(In Chinese)

[18] 尤玉玲,李娟,高孙玉洁,等. ARIMA 模型和 SVM 模型联合在感染性腹泻发病预测中的应用 [J]. 医学动物防制,2020,36(5):432–435.
You YL, Li J, Gao SYJ, et al. Application of ARIMA model and SVM model in the prediction of infectious diarrhea [J]. Journal of Medical Pest Control, 2020, 36(5): 432–435.(In Chinese)

[19] 杨廉平,刘立,刘雨晨,等. 暴雨洪涝影响感染性腹泻发病的环境–社会因素与社会驱动过程模型构建[J]. 环境与职业医学,2022,39(3):296–303.
Yang LP, Liu L, Liu YC, et al. Review on environmental–social factors and social driving process model construction of infectious diarrhea affected by rainstorm and flood [J]. Journal of Environmental & Occupational Medicine, 2022, 39(3): 296–303.(In Chinese)

收稿日期:2025–07–13

读者·作者·编者

温馨提示

为了使杂志能准确投递，敬请作者在单位一栏写明详细科室及联系电话。通信作者必须写明 E–mail 地址。

《现代预防医学》 编辑部