DOI: 10.16835/j.cnki.1000-9817.2025082

· 疾病控制 ·

时间序列和机器学习模型在上海市中小学生因病缺课趋势预测中的应用

王正中，张喆，周欣怡，袁琳琳，翟娅妮，孙力菁，罗春燕

上海市疾病预防控制中心儿童青少年健康所，200336

【摘要】 目的 分析上海市中小学生因病缺课的时间变化规律，探索适用于预测因病缺课高峰时间和强度的模型。**方法** 运用季节趋势分解（STL）法分析上海市中小学生 2010 年 9 月 1 日至 2018 年 6 月 30 日因病缺课率的季节性和长期趋势变化，并采用基于动态时间规整（DTW）的分层聚类方法对时间变化模式相似的缺课症状进行分类。基于历史数据，构建并评估不同时间序列算法和机器学习模型，以优化预测因病缺课趋势的准确性。**结果** 研究期间学年平均新发因病缺课率为 16.86/万人天，缺课趋势呈现季节性和长期上升特征，在 2017 学年达到最高（22.47/万人天）。缺课症状可分为 3 类：冬春高发类（呼吸道症状、发热及全身不适等）、夏季高发类（眼部症状、鼻出血等）以及无明显季节性类（皮肤症状、意外伤害等）。构建的时间序列模型均能较好地预测因病缺课时间趋势，但峰值强度预测的准确度较低，其中多层感知器模型表现最好，均方根误差为 8.96，平均绝对误差为 4.37，相较基准模型分别降低 36.51% 和 39.02%。**结论** 时间序列模型和机器学习算法能够有效预测因病缺课的趋势，在高峰期可针对不同症状导致的缺课采取相应防控措施。

【关键词】 时间；序列分析；疾病；学生

【中图分类号】 R 179　N 945.12　G 478　**【文献标识码】** A　**【文章编号】** 1000-9817（2025）03-0426-05

Application of time series and machine learning models in predicting the trend of sickness absenteeism among primary and secondary school students in Shanghai/*WANG Zhengzhong, ZHANG Zhe, ZHOU Xinyi, YUAN Linlin, ZHAI Yani, SUN Lijing, LUO Chunyan. Institute of Child and Adolescent Health, Shanghai Municipal Center for Disease Control and Prevention, Shanghai* (200336), *China*

【Abstract】 Objective To analyze the temporal variation patterns of sickness absenteeism among primary and secondary school students in Shanghai, so as to explore models suitable for predicting peaks and intensity of absenteeism rates. **Methods** The seasonal and trend decomposition using loess (STL) method was used to analyze the seasonal and long-term trend changes in sickness absenteeism among primary and secondary school students from September 1 in 2010 to June 30 in 2018, in Shanghai. A hierarchical clustering method based on Dynamic Time Warping (DTW) was employed to classify absenteeism symptoms with similar temporal patterns. Based on historical data, the study constructed and evaluated different time series algorithms and machine learning models to optimize the accuracy of predicting the trend of sickness absenteeism. **Results** During the research period, the average new absenteeism rate due to illness was 16.86 per 10 000 person-day for every academic year, and the trend of sickness absenteeism exhibited both seasonality and a long-term upward trend, reaching its highest point in the 2017 academic year (22.47 per 10 000 person-day). The symptoms of absenteeism were divided into three categories: high incidence in winter and spring (respiratory symptoms, fever and general discomfort, etc.), high incidence in summer (eye symptoms, nosebleeds, etc.) and those without obvious seasonality (skin symptoms, accidental injuries, etc.). The constructed time series models effectively predicted the trend of absenteeism due to illness, although the accuracy of predicting peak intensity was relatively low. Among them, the multi-layer perceptron (MLP) model performed the best, with an root mean squared error (RMSE) of 8.96 and an mean absolute error (MAE) of 4.37, reducing 36.51% and 39.02% compared to the baseline model. **Conclusion** Time series models and machine learning algorithms could effectively predict the trend of sickness absenteeism, and corresponding prevention and control measures can be taken for absenteeism caused by different symptoms during peak periods.

【Keywords】 Time; Sequence analysis; Disease; Students

【基金项目】 国家社会科学基金项目（22BTY105）；上海市加强公共卫生体系建设三年行动计划（2023—2025 年）重点学科项目（GWVI-11.1-31）；上海市加强公共卫生体系建设三年行动计划（2023—2025 年）学科带头人项目（GWVI-11.2-XD14）

【作者简介】 王正中（1997- ），女，浙江杭州人，硕士，主要从事学校卫生工作。

【通信作者】 孙力菁，E-mail：sunlijing@scdc.sh.cn；罗春燕，E-mail：luochunyan@scdc.sh.cn

因病缺课监测是了解学生健康和校内传染病流行情况不可或缺的手段[1],国内外学者积极研究其与学校聚集性疫情的关联,构建预测模型预警风险[2]。目前,学校监测系统相对较成熟,但在数据利用与模型构建方面仍不完善,存在数据质量不一、模型验证不足等问题,制约了实际应用效果[3]。本研究基于上海市中小学因病缺课监测的历史数据,全面分析上海地区中小学生的因病缺课规律,探讨导致缺课率在时间尺度上变化的主要原因,同时评估缺课变化规律的可预测性,建立因病缺课的预测模型并评价其适用性,旨在为学校制定科学有效的疾病防控政策提供数据支持。

1　资料来源与方法

1.1　资料来源　研究数据源于上海市疾病预防控制中心的因病缺课网络直报系统,获取 2010 年 9 月 1 日到 2018 年 6 月 30 日期间,共计 8 个学年的上海市全部中小学的因病缺课数据,包括缺课发生日期、缺课性质(新发或既往)等变量。数据导出时,所有学生个人信息(姓名、住址等)均已脱敏,确保不涉及任何个人隐私信息,因此免除伦理学审查。

1.2　相关定义　本研究核心变量为新发因病缺课率,是用以衡量特定周内因新发生的疾病而缺课的指标,新发因病缺课率=(周新发因病缺课例数/周教学总人天数)×10^4[4]。学年新发因病缺课率为周新发因病缺课率按周教学人天数加权的平均值,反映学年的平均缺课情况[5]。症状因病缺课率计算因特定症状造成的缺课发生率。

1.3　统计学分析　数据探索阶段。采用基于局部加权回归的季节趋势分解(seasonal and trend decomposition using loess, STL)对新发因病缺课率序列进行分解,探索其趋势性、季节性和周期性成分。同时,使用动态时间规整(dynamic time warping, DTW)的分层聚类方法,对标准化的症状缺课率进行分析并形成聚类,以探究季节性和周期性存在的潜在原因。

模型构建阶段。将 8 个学年的数据集拆分为训练集和测试集,分别用于构建和评估模型,前者包括 2010 年 9 月至 2017 年 6 月(前 7 个学年)的数据,后者包括 2017 年 9 月至 2018 年 6 月(第 8 个学年)的数据。在训练集上,构建 4 种模型:自回归整合移动平均模型(autoregressive integrated moving average, ARIMA),具有 Box-Cox 变换、ARMA 误差、趋势和季节性分量的指数平滑状态空间模型(trigonometric exponential smoothing state space model with Box-Cox transformation, ARMA errors, and trend and seasonal components, TBATS)模型,支持向量回归模型(support vector regression, SVR)和多层感知器模型(multi-layer perceptron, MLP)[6-10]。其中,ARIMA 和 TBATS 模型是时间序列方法;SVR 和 MLP 模型是机器学习方法。在测试集上计算均方根误差(root mean squared error, RMSE)和平均绝对误差(mean absolute error, MAE)作为评价指标,比较模型性能。同时以基于季节的 Naive 模型(seasonal naive, SNAIVE)为基准,计算构建模型的性能改善率[11]。数据分析使用 R 4.1.0 和 Python 3.10.0 进行。

2　结果

2.1　因病缺课基本情况　2010 年 9 月 1 日到 2018 年 6 月 30 日,上海市中小学因病缺课网络系统累计直报新发因病缺课 3 319 066 人次,平均每学年有 414 883 人次新发因病缺课。学年新发因病缺课率以 2017 学年最高,达 22.47/万人天,8 学年平均新发因病缺课率为 16.86/万人天。

2.2　上海市中小学生新发因病缺课时间序列特征分析　对新发因病缺课时间序列(图 1A)进行 STL 分解,图 1B、C、D 分别为分解后的长期趋势、季节变动和不规则波动部分。由图 1B 可见,新发因病缺课率长期趋势呈波动上升式,其中在 2013—2014, 2017—2018 年有 2 段较快的上升趋势;图 1C 可见,新发因病缺课率季节性模式明显,冬季升高。

时间序列聚类分析结果显示,症状缺课的变化可归纳为 3 种主要模式:A 组症状造成的缺课比例最高(69.37%),主要包括呼吸道症状、发热相关症状和全身不适症状,呈现冬春季节高发特点;B 组症状的发生频率随季节变化不大,主要包括皮肤症状、意外伤害、消化系统症状、心理和学习问题等;C 组症状主要包括眼部症状和鼻出血等,呈现夏季高发特点。见图 2。

2.3　构建模型预测中小学生新发因病缺课情况　经初步探索,针对各模型设定了超参数搜索范围,并用网格搜索法确定最终超参数取值。使用 ARIMA、TBATS、SVR、MLP 模型和基准模型 SNAIVE 对 2017 学年的新发因病缺课率进行预测,由图 3 可见,模型预测值的走势与实际走势大体相符。具体来说,SVR 和 MLP 模型达峰值时间(第 18 周)与真实值同步,但所有模型预测的峰值强度都低于真实值。总体而言,构建的 4 个模型表现均优于 SNAIVE 基准模型,机器学习方法(SVR 和 MLP 模型)在测试集上的表现较传统时间序列方法(ARIMA 和 TBATS 模型)更优。根据 RMSE 指标,各模型的预测性能排序为:MLP > SVR > TBATS > SARIMA > SNAIVE,其中 MLP 模型的预测效果最好,RMSE 为 8.96,MAE 为 4.37,相较于 SNAIVE 模型分别降低 36.51%和 39.02%。见表 1。

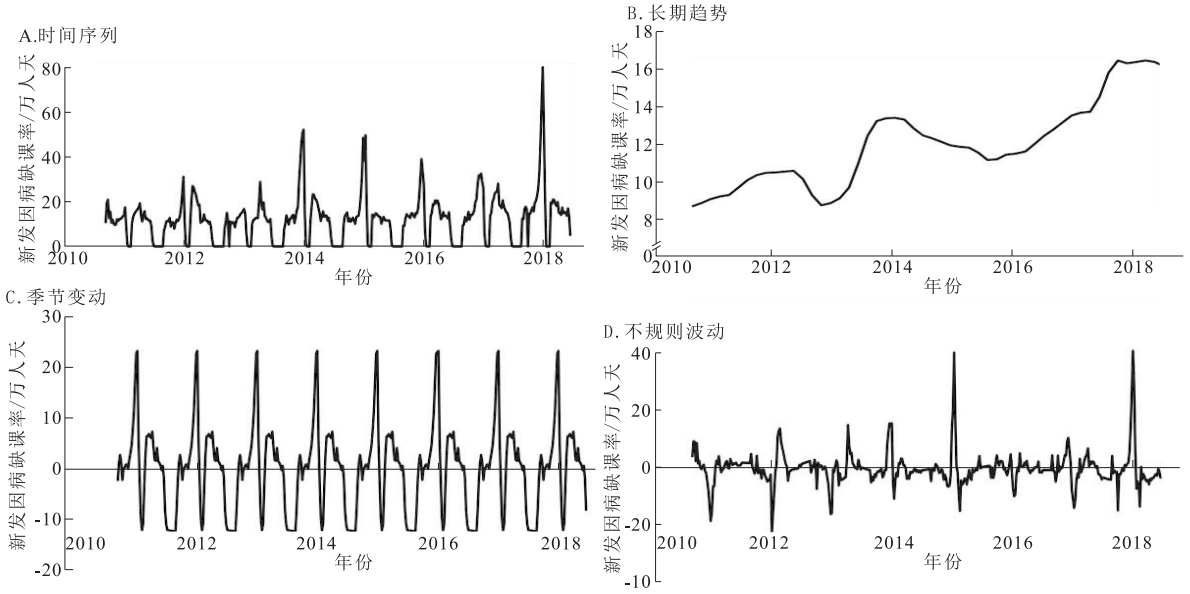


图 1　上海市中小学生新发因病缺课率时间序列季节趋势分解

Figure 1　Seasonal and trend decomposition of the time series of sickness absenteeism rate among primary and secondary students in Shanghai

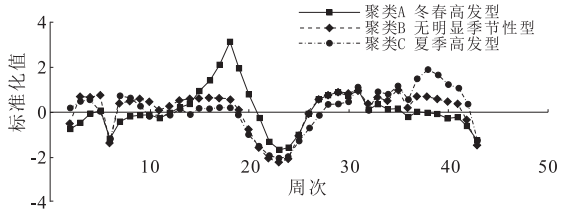


图 2　上海市中小学生不同聚类因病缺课症状季节变化趋势

Figure 2　Seasonal trends in symptoms of sickness absenteeism in different clusters among primary and secondary students in Shanghai

3　讨论

本研究聚类分析结果显示，缺课率变化可归纳为 3 类主要模式，分别为冬春季高发、夏季高发和无明显季节性。建模分析显示，所有模型均能显示比基准模型更优的预测性能，机器学习模型的性能略优于传统时间序列模型，以 MLP 模型的预测准确性最好，但倾向于低估缺课的峰值强度。

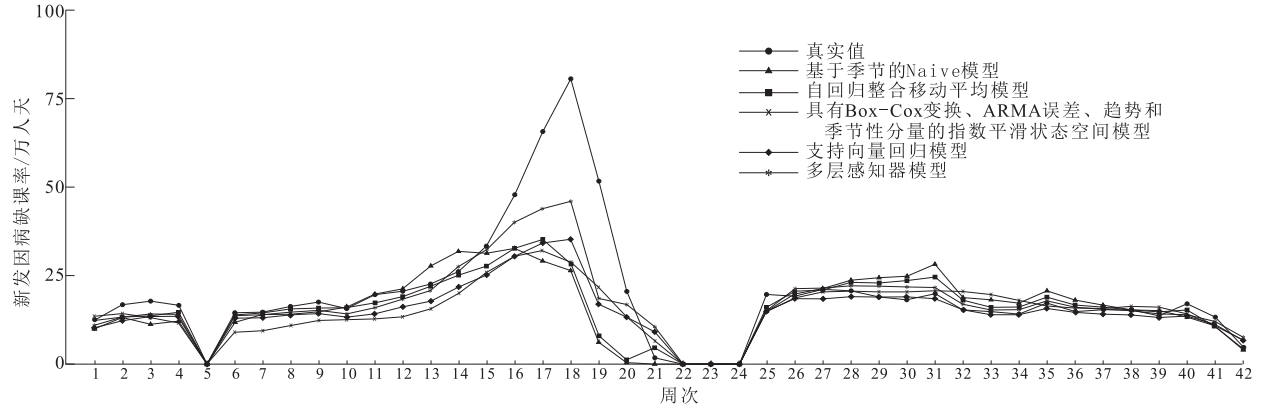


图 3　不同模型预测 2017—2018 学年中小学生新发因病缺课情况

Figure 3　Prediction of sickness absenteeism in 2017-2018 based on different models among primary and secondary students

表 1　不同模型误差指标与改善率的比较

Table 1　Comparison of errors metrics and improvement rates of different models

模型	RMSE	MAE
SNAIVE	14.11(0.00)	7.16(0.00)
SARIMA	12.92(8.40)	5.91(17.54)
TBATS	12.10(14.26)	6.65(7.18)
SVR	11.40(19.18)	5.84(18.52)
MLP	8.96(36.51)	4.37(39.02)

注：( )内数字为改善率/%。

研究结果显示，缺课率呈现季节性变化的主要症状涵盖了呼吸道症状、发热症状以及全身症状，提示其与呼吸道传染病之间存在较强的关联性[12]。相比之下，其他症状如伤害、心理与情绪相关的问题、以及皮肤相关的皮疹等症状，并未显示出明显的季节性变化特征。值得注意的是，诸如鼻出血、结膜红肿、眼眶痛等夏季较为多见的症状，可能与特定疾病（如出血

性结膜炎[13])相关,但在缺课原因中所占比例并不高。

为预防和控制学校传染病,识别学生发病高峰时间和峰值尤为重要[14]。然而,疾病流行情况存在长期变异,因病缺课也逐年变化[15]。为减少这种变化的不确定性带来的影响,可依托预测模型对因病缺课率进行预估。目前,在卫生健康领域应用较多的预测模型主要包括统计学模型(如 ARIMA 模型)和机理模型(如基于传染病动力学的易感-暴露-恢复模型)[16]:前者是数据驱动的,后者则根据疾病发生发展的抽象过程进行建模[17]。但因病缺课事件的发生具有多因性和复杂性,寻找合适的机理建模方法非常困难。因此,本研究在统计学模型中选取了传统时间序列和基于机器学习算法的预测方法。在传统时间序列领域使用了 ARIMA 模型和 TBATS 模型[18],其中 ARIMA 模型是经典方法,具有广泛的应用,而 TBATS 模型能够分析复杂的季节性和周期性。在机器学习领域,存在多类模型适用于处理复杂数据[19]。考虑到样本量的限制以及为了减少过拟合现象,本研究选用了在小样本情况下表现较好的 SVR 模型和 MLP 模型,旨在实现准确性与实用性之间的良好平衡。

传染病发病、传播和流行涉及的因素错综复杂,除季节性外,发病趋势的非线性增加和人群易感性的周期变化都是需要考虑的因素。而传统时间序列模型中平稳性、线性和正态分布假设从一定程度上限制了其捕捉真实世界中复杂的变化规律[4],而机器学习模型能学习数据中的潜在规律,识别复杂的关系,表现出强大的预测性能。在卫生健康领域,机器学习模型的应用已经取得一些进展,其预测结果可作为制定有效防控策略的重要依据[20-21]。本次研究中,所有模型预测的高峰强度都低于实际值,说明仅靠历史数据可能无法准确预测疾病峰值,有存在未被考量的外部因素造成的缺课升高。原因在于病原体、疫苗接种情况的改变可能会引起发病率和缺课率的变化。2017 年冬天流感病毒从往年的甲型为主转变为乙型为主[22],而此前我国多采用的三价流感疫苗,免疫覆盖范围缺乏该季流行的乙型/Yamagata 谱系,也为该流感的流行提供了机会[23]。此外,社会因素也可能起到一定作用,人口密度、公众意识和行为的变化都对总体趋势存在影响。人口密度的增加会促进疾病的传播,从而导致更高的缺课率。同时,家长和老师对学生健康认识和关注的提高也可能促使更多的人选择就医诊疗[24],间接增加了因病缺课的发生。

本文基于学校直报系统的多年历史数据,对缺课高峰模式进行了系统分析,并引入机器学习模型以优化模型构建,提升预测精度,增强了时间尺度上的识别、追溯与预测能力。鉴于目前全国大多数地区已普遍采用电子化平台收集学生因病缺课缺勤信息,通过合理设定参数范围并应用高效预测模型,系统能够在后台自动调整相应参数,并及时生成预警信息,不仅为学校疾病防控提供了更为准确的数据支持,还为制定防控策略提供了科学依据。根据模型前瞻性预测结果,学校卫生相关部门可优化资源配置,加强高峰期监测力度,增加卫生物资储备。同时,学校也可以根据模型预测的情况加强晨检和每日健康巡检,必要时增加午检和晚检等措施,并开展宣传活动,以加强高峰期的监测和个人防护。然而,该模型在评估疫情高峰强度时表现不佳,应用存在一定局限。为提高模型实用性,可以考虑整合其他数据源,如互联网搜索数据、大数据中心的学生临床就诊数据等[12]。通过综合个体或区域层面信息,更准确地评估因特定原因引起的缺课,发现潜在流行病聚集情况。另外,实际应用中,需密切关注预测模型表现,在特殊条件下,比如采取大规模线上教学、停课等应急措施时[25],模型的预测结果可能会受到影响,需要不断进行调整和优化。

本研究深入探讨了因病缺课率变化的趋势,并揭示了呼吸道症状、发热症状以及全身症状是缺课率变化的主要因素。同时通过评估多种模型对缺课率趋势的预测能力,发现 MLP 模型和 SVR 模型在估计因病缺课高峰时间方面表现更佳。对于因病缺课高峰强度的预测,仍需结合外部因素进行考量。

利益冲突声明　所有作者声明无利益冲突。

参考文献

[1] 马军. 中国学校卫生/儿少卫生发展[J]. 中国学校卫生, 2015, 36(1): 6-9.
MA J. School health/ child and adolescent health development in China[J]. Chin J Sch Health, 2015, 36 (1): 6-9. (in Chinese)

[2] TSANG T K, HUANG X, GUO Y, et al. Monitoring school absenteeism for influenza-like illness surveillance: systematic review and Meta-analysis[J]. JMIR Public Health Surveill, 2023, 9: e41329.

[3] DONALDSON A, HARDSTAFF J, HARRIS J, et al. School-based surveillance of acute infectious disease in children: a systematic review[J]. BMC Infect Dis, 2021, 21(1): 744.

[4] SCHMIDT W P, PEBODY R, MANGTANI P .School absence data for influenza surveillance: a pilot study in the United Kingdom[J]. Eur Surveill, 2010, 15(3): 19467.

[5] 郝莉, 朱冰, 施文英, 等. 2013—2017 学年浙江省杭州市中小学生因病缺课监测分析[J]. 疾病监测, 2020, 35(9): 840-844.
HAO L, ZHU B, SHI W Y, et al. Surveillance for illness-induced

school absence in students of elementary and secondary schools in Hangzhou, 2013-2017[J]. Dis Surveill, 2020, 35(9): 840-844. (in Chinese)

[6] BOX G E P, JENKINS G M, REINSEL G C, et al. Time series analysis: forecasting and control[M]. Hoboken, New Jersey: John Wiley & Sons Inc, 2016: 88-106.

[7] HYNDMAN R J, KHANDAKAR Y. Automatic time series forecasting: the forecast package for R[J]. J Stat Softw, 2008, 27(3): 1-22.

[8] DE LIVERA A M, HYNDMAN R J, SNYDER R D. Forecasting time series with complex seasonal patterns using exponential smoothing[J]. J Am Stat Assoc, 2011, 106(496): 1513-1527.

[9] SMOLA A J, SCHÖLKOPF B. A tutorial on support vector regression [J]. Stat Comput, 2004, 14(3): 199-222.

[10] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors [J]. Nature, 1986, 323 (6088): 533-536.

[11] SVETUNKOV I, PETROPOULOS F. Old dog, new tricks: a modelling view of simple moving averages[J]. Int J Prod Res, 2018, 56 (18): 6034-6047.

[12] KARA E O, ELLIOT A J, BAGNALL H, et al. Absenteeism in schools during the 2009 influenza A(H1N1) pandemic: a useful tool for early detection of influenza activity in the community? [J]. Epidemiol Infect, 2012, 140(7): 1328-1336.

[13] 胡停停, 赵鹤鹤, 段晓健, 等. 2013—2020 年我国急性出血性结膜炎流行特征及暴发疫情分析[J]. 疾病监测, 2021, 36(5): 440-444.
HU T T, ZHAO H H, DUAN X J, et al. Epidemiological characteristics and outbreak analysis of acute hemorrhagic conjunctivitis in China, 2013-2020[J]. Dis Surveill, 2021, 36(5): 440-444. (in Chinese)

[14] LUTZ C S, HUYNH M P, SCHROEDER M, et al. Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples[J]. BMC Public Health, 2019, 19(1): 1659.

[15] 孙婕, 杨雯雯, 曾令佳, 等. 2011—2016 年全国 6~22 岁学生人群法定传染病监测数据分析[J]. 中华流行病学杂志, 2018, 39 (12): 1589-1595.
SUN J, YANG W W, ZENG L J, et al. Surveillance data on notifiable infectious diseases among students aged 6-22 years in China, 2011-2016[J]. Chin J Epidemiol, 2018, 39 (12): 1589-1595. (in Chinese)

[16] 黄硕, 刘才兄, 邓源, 等. 世界主要国家和地区传染病监测预警实践进展[J]. 中华流行病学杂志, 2022, 43(4): 591-597.
HUANG S, LIU C X, DENG Y, et al. Progress in the practice of surveillance and early warning of infectious diseases in major countries and regions[J]. Chin J Epidemiol, 2022, 43 (4): 591-597. (in Chinese)

[17] CHOWELL G, SATTENSPIEL L, BANSAL S, et al. Mathematical models to characterize early epidemic growth: a review[J]. Phys Life Rev, 2016, 18: 66-97.

[18] 沈忠周, 马帅, 曲翌敏, 等. ARIMA 模型在我国法定传染病报告数中的应用[J]. 中华流行病学杂志, 2017, 38(12): 1708-1712.
SHEN Z Z, MA S, QU Y M, et al. Application of autoregressive integrated moving average model in predicting the reported notifiable communicable diseases in China[J]. Chin J Epidemiol, 2017, 38 (12): 1708-1712. (in Chinese)

[19] PETROPOULOS F, APILETTI D, ASSIMAKOPOULOS V, et al. Forecasting: theory and practice[J]. Int J Forecast, 2022, 38(3): 705-871.

[20] 付之鸥, 周扬, 陈诚, 等. 时间序列分析与机器学习方法在预测肺结核发病趋势中的应用[J]. 中国卫生统计, 2020, 37(2): 190-195.
FU Z O, ZHOU Y, CHEN C, et al. Application of time series analysis and machine learning methods in predicting the incidence of tuberculosis[J]. Chin J Health Stat, 2020, 37 (2): 190-195. (in Chinese)

[21] 杨慧, 魏麟, 胡晓斌, 等. 基于 CEEMD-GWO-SVR 集成模型的病毒性肝炎流行趋势预测研究[J]. 中国卫生统计, 2022, 39 (6): 815-818.
YANG H, WEI L, HU X B, et al. Epidemiological trend prediction of viral hepatitis based on CEEMD-GWO-SVR integration method [J]. Chin J Health Stat, 2022, 39 (6): 815-818. (in Chinese)

[22] 孔德川, 吴寰宇, 郑雅旭, 等. 上海市 2015—2017 年成年人急性呼吸道感染病例的流行病学和病原学特征分析[J]. 中华流行病学杂志, 2019, 40(8): 904-910.
KONG D C, WU H Y, ZHENG Y X, et al. Etiologic and epidemiologic features of acute respiratory infections in adults from Shanghai, during 2015-2017[J]. Chin J Epidemiol, 2019, 40 (8): 904-910. (in Chinese)

[23] YAO Y, CHEN L, ZHU D, et al. Increasing serum antibodies against type B influenza virus in 2017-2018 winter in Beijing, China[J]. AMB Express, 2022, 12(1): 127.

[24] 张晓波,施鹏,郑珊, 等. 上海市单中心儿科门诊特征和医疗服务 2009—2018 年趋势分析[J]. 中国循证儿科杂志, 2019, 14(3): 161-168.
ZHANG X B, SHI P, ZHENG S, et al. An analysis of characteristics and trends of outpatient visits and medical services in a tertiary pediatric hospital in Shanghai from 2009 to 2018[J]. Chin J Evid-Based Pediatr, 2019, 14(3): 161-168. (in Chinese)

[25] 张欣. 新发传染病是学校卫生面临的永恒挑战[J].中国学校卫生,2020, 41(6): 641-644.
ZHANG X. Perpetual challenges of emerging infectious disease of school children[J]. Chin J Sch Health, 2020, 41 (6): 641-644. (in Chinese)

收稿日期:2024-04-02 **修回日期**:2024-12-04 **本文编辑**:孙曼莉