

基于元胞自动机模型模拟松材线虫病的传播

周宏威, 李永正, 陈怡帆, 胡浩昌, 张思岩, 崔迪

(东北林业大学 黑龙江 哈尔滨 150040)

摘要:【目的】在之前的松材线虫病预测研究中, 使用的影响因素往往不够全面, 只考虑了自然因素的作用, 缺少对人为影响因素的利用, 同时在建模过程中忽略了传播扩散过程和对空间信息的利用。为了探究影响松材线虫病传播扩散的主要影响因素, 结合人类活动因素和空间特征, 围绕松材线虫“传入-定殖-扩散”的生态入侵过程, 构建一种适用于更小空间尺度数据的预测模型, 实现对高风险发生地区的精准预测和早期预警。【方法】本研究使用国家林业草原局获得的江苏省松材线虫病小班本底历史发生数据, 基于松材线虫病的生态和地理分布特点, 选取自然气候、地理信息以及人类活动因素作为影响因子数据, 通过主成分分析技术 (PCA) 进行预处理, 结合贝叶斯估计方法进行对影响因子数据进行特征增强, 探究不同影响因子对于松材线虫病发生的作用关系。建立灰狼优化算法—元胞自动机模型来模拟松材线虫病的入侵扩散过程。同时与支持向量机模型 (SVM), 随机森林模型 (RF), 分类与回归树 (CART), 广义线性模型 (GLM), 人工神经网络模型 (ANN) 五种主流机器学习模型预测结果进行对比, 对江苏省 2022 至 2023 的松材线虫病发生情况进行了模拟, 计算召回率和精确率等评价指标对模型的性能进行比较。【结果】本研究所构建的灰狼优化算法—元胞自动机模型的召回率达到了 80.4%, 明显优于其他五种机器学习模型, 同时 F1 分数达到了 45.3%, 显示模型在保持预测结果准确性的同时, 还能有效地对松材线虫病新发小班进行预测。通过对比分析, 本研究不仅还强调空间特征在松材线虫病传播预测研究中的重要性, 同时证实元胞自动机模型在处理复杂时空数据和更精细尺度空间数据预测方面的高度适用性。【结论】本研究证实了温度、降水以及人口和 GDP 是影响松材线虫病传播扩散的主要因素, 松材线虫的发生主要集中在低海拔的地区。元胞自动机模型作为一种结合空间信息和时间序列的模型, 在处理复杂的生态数据和入侵物种风险评估中具有更高的适用性, 元胞自动机模型可以作为一种有效的工具为精准高效地采取预防措施提供技术支持。

关键词: 松材线虫病; 传播预测模型; 贝叶斯估计; 元胞自动机模型; 灰狼优化算法

中图分类号 Q958.15; S154.5

文献标识码: A

文章编号: 1001-7488 ()

-0000-00

Simulating the spread of pine wilt disease based on cellular automaton model

Abstract: 【Objective】Previous studies on the prediction of pine wood nematode (PWN) disease often lacked comprehensiveness, focusing solely on natural factors and neglecting human influences and the process of spread and diffusion. In response, this study aims to investigate the major factors influencing the spread of PWN by integrating human activity and spatial characteristics. Centered around the ecological invasion process of "introduction-settlement-diffusion," a predictive model suitable for smaller spatial scale data was developed to enable precise forecasting and early warning in high-risk areas. 【Method】This study

utilized historical incidence data of PWN in Jiangsu Province obtained from the National Forestry and Grassland Administration. Based on the ecological and geographical distribution characteristics of the disease, natural climate, geographical information, and human activity factors were selected as the data variables. The data was preprocessed using Principal Component Analysis (PCA) to reduce dimensions and extract significant features, coupled with Bayesian estimation methods to enhance the features of the impact factors. A Grey Wolf Optimizer–Cellular Automata model was developed to simulate the invasion and spread of PWN. This model was compared against five mainstream machine learning models: Support Vector Machine (SVM), Random Forest (RF), Classification and Regression Trees (CART), Generalized Linear Model (GLM), and Artificial Neural Network (ANN). The models were used to simulate the PWN occurrence in Jiangsu Province for the years 2022 to 2023, evaluating their performance through metrics such as recall and precision rates. 【Result】The Grey Wolf Optimizer–Cellular Automata model achieved a recall rate of 80.4%, significantly outperforming the other models, with an F1 score of 45.3%. This demonstrates the model's accuracy in predicting new occurrences of PWN while maintaining high predictive precision. The study not only highlights the importance of spatial features in predicting the spread of PWN but also confirms the high applicability of the Cellular Automata model in managing complex spatio-temporal data and fine-scale spatial predictions. 【Conclusion】This study confirms that temperature, precipitation, population, and GDP are primary factors influencing the spread of PWN, with occurrences mainly concentrated in low-altitude areas. The Cellular Automata model, integrating spatial information and time series data, proves highly suitable for handling complex ecological data and assessing risks of invasive species. As an effective tool, the Cellular Automata model provides technical support for taking precise and efficient preventative measures against PWN disease.

Key words: Pine Wilt Disease; Spread Prediction Model; Bayesian Estimation; Cellular Automaton Model; Grey Wolf Optimizer

由于气候变化和人类活动频繁,伴随国际贸易而来外来物种入侵已经成为各个国家面临的重要生态安全问题之一。松材线虫病是由松材线虫(*Bursaphelenchus xylophilus*)引起的一种毁灭性的森林病害,又被称为松树的“癌症”,是我国目前最为严重的林业灾害之一^[1]。自1982年传入中国以来,已迅速扩散至全国20多个省(区、市),920个县级行政区,对中国南方数百万公顷的松林构成毁灭性威胁,导致大量松树枯死,造成了巨大的经济损失,严重威胁我国的森林生态系统^[2]。

机器学习技术可以从训练数据出发,寻找数据之间复杂的规律,在分类和回归任务中都表现出了广泛的适用性,例如CART^[3]、随机森林模型^[4]、支持向量机模型^[5]等都被广泛用于松材线虫病传播预测研究中。Zhang和Dian等人结合了人为活动,地形变量,包括海拔、坡度和坡向等因素,比较了随机森林(RF)、K邻域(KNN)、支持向量机(SVM)和人工神经网络(ANN)等不同机器学习算法的拟合效应并预测了中国地区松材线虫病的发生概率^[6]。

物种分布模型(Species Distribution Model, SDM)也称生态位模型(Ecological Niche Models, ENM)被广泛用于松材线虫适生区预测研究中^[7-8]。该类模型在环境因子组成的多维生态空间中,依据采样点提供的统计信息估计物种的生态位需求,然后投射到选定的时空范围内,以概率的形式反映物种对生境的偏好程度,模型结果通常反映大尺度上物种适宜生境分布^[9]。

2021年, Tang 和 Yuan 等人采用 MaxEnt 模型确定了中国松材线虫的当前和潜在分布及生境适宜性, 使用 2050 年和 2070 年的气候数据, 预测了未来两种不同的气候变化情景下, 松材线虫并在中国的潜在分布^[10]。

但是在模型应用过程中, 研究者往往对所研究的生态问题缺乏清楚的认识与表达, 对模型算法和参数并不了解, 仅仅通过文献模仿前人的工作, 且没有针对性的筛选建模数据与模型算法, 导致不确定性较大且难以解释的模型结果^[11], 此外只使用气候变化因素对于未来进行长时间、大尺度的预测, 预测结果的准确性和适用性还需要进一步验证。

综合之前的研究, 还存在以下的缺陷: 首先, 在之前的松材线虫病预测研究中大多数只考虑了自然因素的作用, 使用物种生态位模型或者机器学习算法, 缺乏对人为活动因素等信息的利用, 忽略了对入侵扩散过程的考虑。最新的研究中也指明, 松材线虫病的扩散在很大程度上是因为人为运输疫木及其制品的原因造成的^[12]。因此在预测过程中, 纳入人类活动的影响, 考虑地区之间的空间连接关系和疫情传播的方向性, 对提高预测精度有重要的研究意义。

其次, 无论是生态位模型或者机器学习模型都是一种复杂的黑盒模型, 我们无法观测模型内部的计算过程从而判断输出结果的准确性并且不能直接对模型的计算方式或者运行参数进行设计, 但模型表现出较差的性能时, 难以对模型进行优化。

最后, 在之前的研究中大多使用县区尺度数据进行预测, 在空间尺度上难以满足精准防控的需要, 在未来, 应该采用更高分辨率、更小空间尺度下的数据集来进行松材线虫病的预测研究, 同时由于小班的空间分布情况复杂, 这对模型的适用性和性能都提出了更高的要求, 对于模型所需要的影响因子需要进行更深层次的挖掘和处理。因此亟待研究出一种有效的方法, 模拟松材线虫的生态入侵过程, 实现高风险发生地区监测预警, 为采取预防措施提供理论指导。

近年来, 元胞自动机 (cellular automata, CA) 模型因其直观性、灵活性、开放结构以及过程中整合空间和时间维度的能力而被广泛用于模拟城市增长模式和预测未来的土地利用变化^[13]。许多学者已经成功地开发和应用了各种基于 CA 的模型, 文献^[14]将 CA 与 SEIR 模型结合, 构建了新型冠状病毒肺炎传播预测模型。2022 年, Liu 和 Zhang 等人将 CA-MA 模型应用于入侵物种的预测研究中, 采用空间自相关分析、方向分析等空间分析方法, 分析了 2000 年、2010 年、2020 年和 2030 年松材线虫发生数据的变化趋势, 揭示了安徽省松材线虫灾害的传播规律^[15]。综合以上的优点, 元胞自动机模型在空间分析过程中具有很大的优势, 在城市增长、生态系统模拟、疾病传播等众多领域得到了广泛应用。

本文使用元胞自动机模型, 通过结合时间序列数据和空间数据, 从而更准确地预测松材线虫的潜在发生区域。首先采集自然因素、人为活动、地理信息等影响因素, 使用核密度分析和贝叶斯方法, 对每种影响因子进行处理, 计算归一化参数作为模型的输入。使用灰狼优化算法—元胞自动机模型, 计算空间相关参数, 以江苏省为例, 对松材线虫病小班尺度数据下未来传播扩散趋势进行预测。元胞自动机模型的引入可以基于历史发生数据和扩散规律建立来模拟入侵物种在局部空间内的扩散趋势, 并结合生态需求, 准确预测未来的发生发展趋势。这种模型的引入, 将为传统的预测模型增添新的研究方法和更深层次的解释能力, 为指导入侵物种的防控措施提供重要理论支持和指导意义。

1 构建数据库

1.1 数据采集与处理

1.1.1 松材线虫病小班数据

松材线虫病的历史发生数据来自于国家森林草原局 (NFGA), 研究区域为江苏省, 时

间跨度为 2021 年至 2023 年。数据集包含松材线虫病在最小森林单元小班中的疫情发生状况（图 2），并且对疫情发生状态缺失或者异常值进行了处理。同时我们获取了存储小班地理信息的矢量格式数据，以便于在地理信息系统对影响因子数据进行关联和处理。

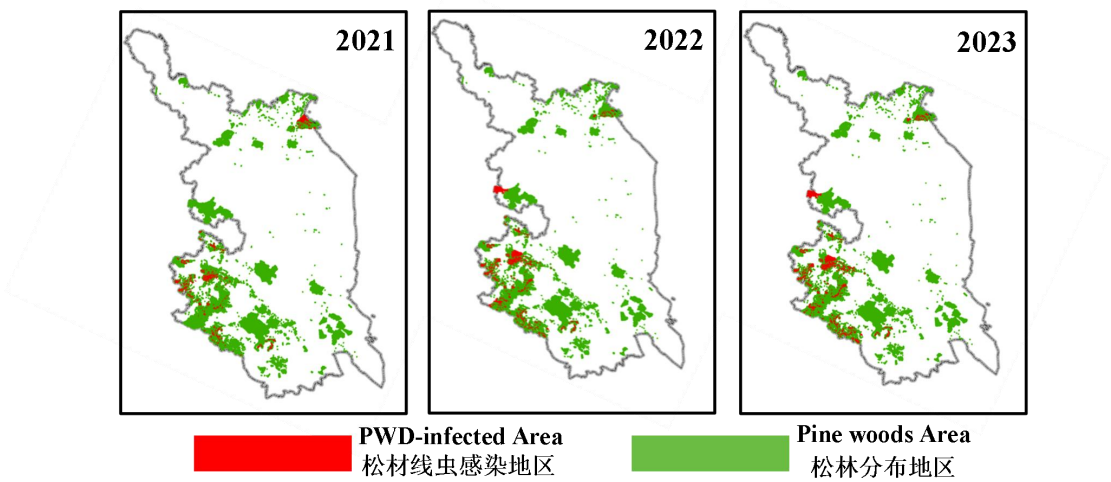


图 2 江苏省松材线虫病发生分布

Fig. 2 Distribution of PWD in Jiangsu Province

1.1.2 影响因子数据

根据已有研究和文献表明自然因素在短距离传播中占据主导地位，降水量和温度是影响松材线虫病发生的主要自然因素^[16-17]，本文选取了包括年均温度、年均降水量、海拔等在内的 19 项物候因子和 1 个地理信息数据以及人口、GDP、道路密度、木材运输数据（表 1）。数据来源于全球气候数据库 2.1 版本（worldclim，<https://www.worldclim.org/>），数据类型为 tif 格式，空间分辨率为 30”，采用 WGS_84 地理坐标系，在 Arcgis 软件中，使用江苏省界进行分割，按照均值方法提取至每个小班中并存储。

人口和 GDP 分布数据来自全球变化研究数据发布与储存库（<http://www.geodoi.ac.cn/>），采集时间为 2020 年，数据类型为 tif 格式，在 Arcgis 软件中使用相同的方法至每个小班中并存储。

道路数据道路分布数据来源于北京大学城市与环境学院地理数据平台（<http://geodata.pku.edu.cn>），数据格式为 shapefile，包括县道、乡镇道路和三级、四级道路，使用 Python 中 Geopandas 软件包读取小班地图数据和道路图数据，通过小班内不同分级道路的总长度加权除以小班面积计算。木材运输数据来自每个县区的木材运输记录，并按照权重分配给每个小班。1.1.3 数据预处理

本文选取了包括年平均温度、降水量在内的 19 项物候因子充分温度和降水对松材线虫及传媒昆虫生长发育的影响及生存和传播的影响，同时 19 项物候因子之间也必然会出现高度的相关性，造成输入数据的冗余，高维度复杂的训练数据不利于模型的拟合并且会影响模型的性能。

为了确保输入变量之间相互独立，去除数据噪声，降低算法计算运算成本，在之前的研究中多使用主成分分析（PCA）方法^[18]或者运用统计分析系统（SAS）^[19]对数据进行相关性分析和预处理工作。

表 1 影响因子数据

Tab. 1 Environment variables data

影响因子类型 Environmental factor type	影响因子 Environmental factor	单位 Unit	变量代称 Code name
生物气候因子 Bioclimatic factor	年均温度 Annual mean temperature	°C	Bio1
	平均温度日较差 Mean diurnal range	°C	Bio2
	等温性 Isothermality	%	Bio3
	温度变化方差 Temperature seasonality	%	Bio4
	最热月极端高温 Max temperature of warmest month	°C	Bio5
	最冷月极端低温 Min temperature of coldest month	°C	Bio6
	年温变化范围 Temperature annual range	°C	Bio7
	最湿季平均温 Mean temperature of wettest quarter	°C	Bio8
	最干季平均温 Mean temperature of driest quarter	°C	Bio9
	最暖季平均温 Mean temperature of warmest quarter	°C	Bio10
	最冷季平均温 Mean temperature of coldest quarter	°C	Bio11
	年均降水量 Annual precipitation	mm	Bio12
	最湿月降水量 Precipitation of wettest month	mm	Bio13
	最干月降水量 Precipitation of driest month	mm	Bio14
	降水量变化方差 Precipitation seasonality	%	Bio15
	最湿季降水量 Precipitation of wettest quarter	mm	Bio16
	最干季降水量 Precipitation of driest quarter	mm	Bio17
	最暖季平均降水量 Precipitation of warmest quarter	mm	Bio18
	最冷季平均降水量 Precipitation of coldest quarter	mm	Bio19
地形因子 Terrain factor	海拔 Elevation	m	Ele
人为影响因子 Human influence	人口 Poplutaion	人	Pop
	地区生产总值 GDP	万元	GDP
	道路密度 Road densit	/Km	R.dens
	木材调运 Timber transport	m ³	T.trans

PCA 是一种基于从高维空间映射到低维空间的映射方法,其目标是向数据变化最大的方向投影,向重构误差最小化的方向投影,能够有效地降低数据维度,简化数据的复杂性,同时保留数据的主要特征,并且能够有效地去除数据中的噪声和异常值^[20]。此外,PCA 将数据投影后得到的特征项链具有更直观的含义,结果更具有解释性,是一种主流的数据分析和处理方法。

本文使用 PCA 数据降维方法,对 Bio1 至 Bio11 和 Bio12 至 Bio19 气温和降水两类影响因子分别进行降维处理,得到两个一维的合成向量分别代替气温类和降水类的影响因子数据用于后续的分析 and 模型训练过程。基于数据预处理结果,使用 Spearman 相关性分析方法,对所有影响因子与发生情况进行相关性检验(图 3)。

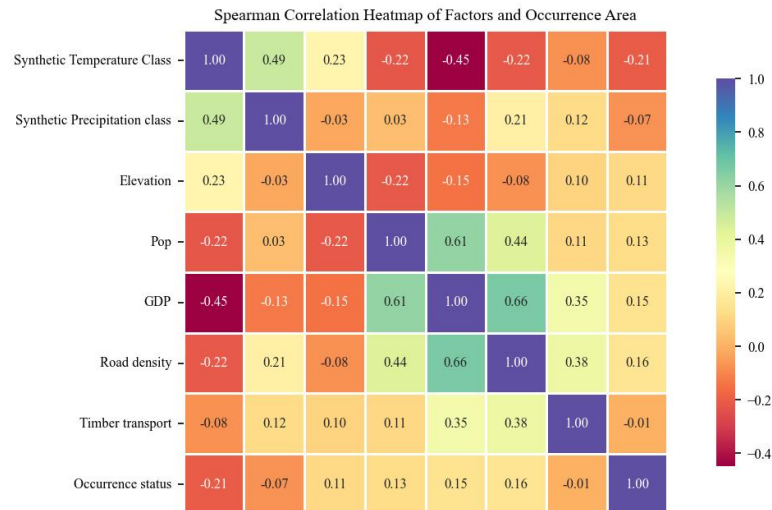


图 3 Spearman 相关性系数矩阵

Fig. 3 Spearman correlation coefficient matrix

可以看出,不同影响因子之间并没有表现出较高的冗余性,同时,为了评估每种因子对结果的影响百分比,使用决策树的特征重要性 (Feature Importance) ^[21]来评估每种因子对于疫情发生所占重要性的百分比 (图 4)。

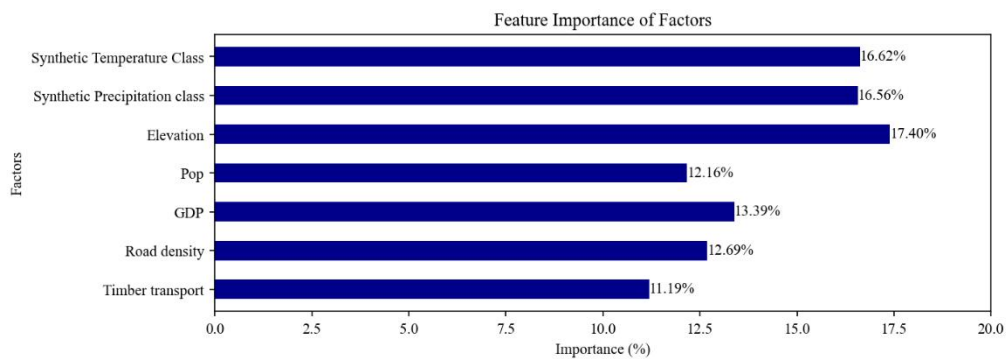


图 4 影响因子重要性分析

Fig. 4 Importance analysis of influencing factors

从相关性和重要性两种分析结果来看,自然因子即通过 PCA 降维处理合成后的气温类和降水类影响因子和疫情发生之间都表示出较高的相关性。而对于人为影响因子来说,人口和 GDP 表现出了更高的重要性。但是由于小班数据集中负样本即不发生的小班地区总量较大和正常本即疫情发生地区的数量极为不平衡,如果只采用归一化、标准化对数据进行放缩处理,势必会影响模型的性能,但是又不适合采用数据增强 (Data Augmentation) 的方法来通过生成新的样本来扩充数据集。需要使用数据挖掘或者特征工程的方法进行处理,使得模型能够更好地拟合影响因子与发生情况之间的关系。

2 方法与结果

2.1 贝叶斯估计

贝叶斯估计是统计学领域中一个非常重要的领域,它利用了先验知识和观测数据来计算后验分布概率,从而对参数进行估计,它关注的是样本数据的分布情况,可以有效地处理由于不确定性和随机性导致真实数据采集中的误差。贝叶斯估计或者极大似然估计都被广泛用于数据预测或者数据预处理过程中或者与其他模型进行集成^[22]。

松材线虫病的发生是由于多种影响因子共同作用的结果, 利于贝叶斯估计方法可以计算出多种影响因子的作用相互叠加的联合概率密度分布函数, 但是也势必会需要大量的计算时间, 参数估计结果的正确性也难以进行验证。因此在数据预处理过程中, 只对单一影响因子和发生情况之间进行参数计算, 基于核密度估计方法分别估计每个特征点在疫情发生和不发生条件下和的概率密度, 通过贝叶斯估计, 计算后验概率作为模型的输入, 计算公式如下:

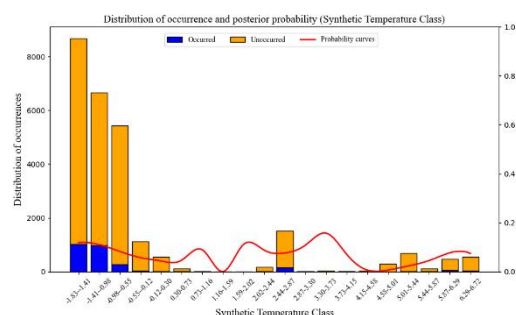
$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B|A) \times P(A) + P(B|A^c) \times P(A^c)}$$

其中, A 代表该小班发生松材线虫病, B 代表该小班中影响因子特征值。 $P(A|B)$ 代表根据不同影响因子计算出的发生松材线虫病的后验概率, $P(B|A)$ 代表每个地区影响因子发生疫情的概率密度。 $P(A)$ 表示疫情发生的先验概率。 $P(A^c)$ 表示不发生松材线虫病, 最终计算通过以上似然概率和先验概率来计算, $P(B|A)$ 即根据每个地区的影响因子来估计发生松材线虫病的后验概率。

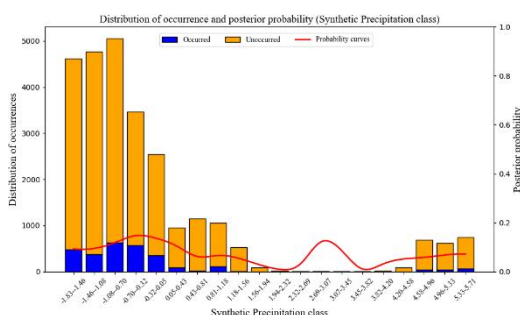
利用贝叶斯定理计算特征与目标变量之间的后验概率, 并将这些概率作为新的特征输入到模型中, 这种方法有助于捕捉特征和目标变量之间的统计关系, 有助于提升模型的分类性能, 增强模型的表现。

2.1.1 影响因子分析

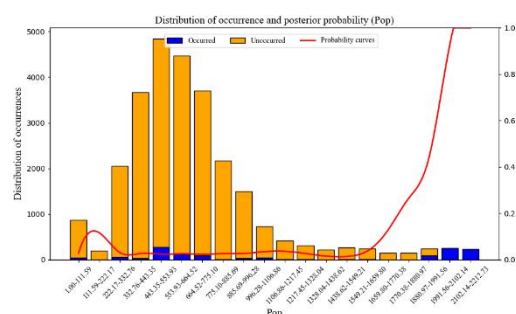
在 Python 中使用 sklearn—KernelDensity 软件包完成对每种影响因子不同特征值下的核密度估计和后验概率计算, 核密度估计过程中, 具体参数设置为 `kernel=gaussian`, 使用高斯核函数, `bandwidth`: 带宽参数从 0.1 到 1.0 之间, 步长为 0.045 (共 20 个候选值), `cv=5` 使用 5 折交叉验证来选择最佳带宽。对结果进行可视化 (图 5), 图中 a—f 分别表示不同影响因子与疫情发生之间的数量统计和概率密度分布函数。



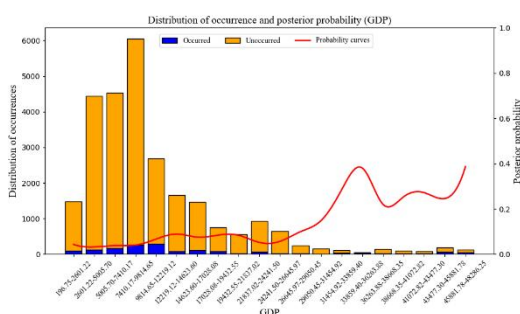
a 气温类因子合成



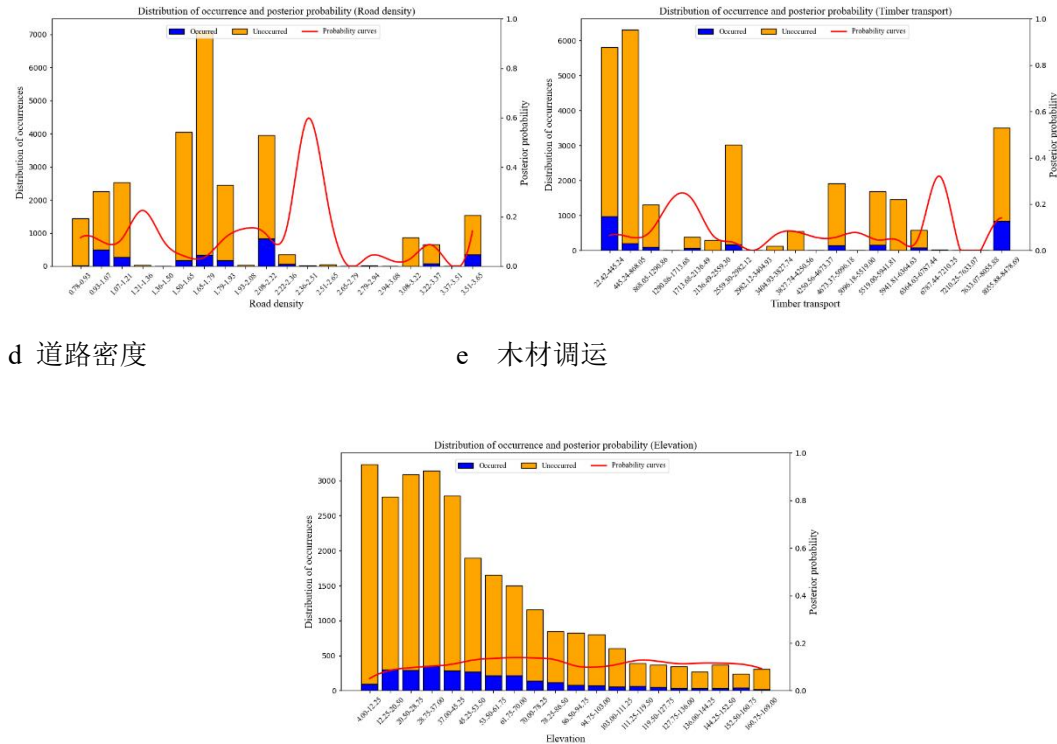
b 降水类因子合成



c 人口



d GDP



d 道路密度

e 木材调运

f 海拔

图 5 影响因子发生情况统计与概率密度分布

Fig. 5 Statistics on the occurrence of influencing factors and probability density distribution

图中蓝色代表发生松材线虫病地区的在该影响因子区间内的数量,黄色代表未发生地区的数量,红色代表通过贝叶斯估计方法拟合得到的发生概率密度估计函数曲线。

图中 a 和 b 展示了松材线虫病发生情况在温度和降水两种影响因子下的分布情况,可以看出松材线虫病的发生地区存在明显的集中情况成两侧集中分布,在气温和降水与松材线虫病发生之间存在重要的相关性。f 展示了在不同海拔下的分布情况,低海拔地区是松材线虫病的高风险发生地区,印证了文献^[23]的结论。

图中 c-e 表示四种人类活动因素对松材线虫病发生的影响。发生地区主要集中在人口数量 332-775 和 1770-2212 两个区间内,同时集中在 196-9814 较低 GDP 的区间内,可能说明较低的经济发展水平的地区对木材的存在更高的需求量而导致松材线虫病的传入和发生。对于道路密度和松材调运的统计分析情况,病害的发生没有显示出显著的集中分布或者单调关系,由此也可以说明,如果直接将这些类型的数据输入到模型中,可能会影响模型的性能,通过核密度分析以及贝叶斯方法进行后验概率估计是有必要的。

2.2 模型建立

元胞自动机模型在模拟复杂系统动态行为方面具有独特的优势,特别是在模拟入侵物种的地理扩散过程中。本研究将通过元胞自动机模型来模拟松材线虫在不同环境条件下的历史扩散过程,结合自然生态因素和人为干预的影响,来预测未来江苏省松材线虫传播扩散趋势。

目前已经开发了 CA-Markov^[24], ANN-CA^[25]等多种 CA 与其他方法集成的软件来进行元胞自动机的建模工作,但是利用已有的集成软件更适合于长时间跨度、高分辨率的数据集,对于 2020 至 2022 的松材线虫小班数据,也可以通过设定转换规则的方式实现元胞自动机模型的建模过程。元胞自动机是一种空间相互作用和时间因果关系为局部的网格动力学模型,主要由四个元素组成:像元、像元状态、邻域和变换规则,其中的核心是过渡规则,文献^[26]

给出了基于单元的元胞自动机模型中过渡规则的一般计算方法,同时也根据已有数据特点进行对其进行修改:

$$p_k = \frac{1}{1 + \exp(-z_k)} \times \text{constraint}_k \times (1 + (-\ln\gamma)) \times \Omega_k$$

其中 p_k 是单元 K 的感染风险; constraint_k 是单元格 K 的约束分数,表示各种影响因子的相互作用关系。 γ 是一个随机数,范围从 1 到 e ,用与描述松材线虫传播扩散的不确定性; z_k 是单元 K 与其他单元格的空間联系强度, Ω_k 表示单元格 K 的邻域被感染程度。

$$\text{constraint}_k = \prod_{n=1}^i (w_n \times b_{i,n} + x_n)$$

其中, $b_{n,k}$ 表示单元 K 的第 n 个影响因子; w_n 是第 n 个影响因子的可调节权重参数, x_n 也表示一个可调节参数。

在 z_k 时计算由于小班的形状不规则并且空間上的分布是离散的,难以对其进行网格划分。文献^[27]提出了一种时空卷积网络深度学习框架,在处理空間信息时,不采用网格或者分段的方式进行处理,而将交通网络建模为节点连接矩阵来描述空間上的连接关系。对与两个任意节点之间的连接可以采用以下的计算方式:

$$z_k = \sum_i^n \left(\frac{S_{j,i} \times \text{GDP}_i \times \text{Pop}_i}{d_{ij}^2} \right)_{\text{normalized}}$$

在计算这些连接关系时,使用 j 单元和 i 单元之间的距离 $s_{j,i}$ 比上距离的平方 d_{ij}^2 ,来计算两个单元之间的空間联系强度。在实际计算过程中,只考虑了疫情发生单元 i 对于单元 i 的影响,同时对每个单元 i 来说,对所有来自其他感染单元 i 的空間联系强度进行累加求和,以计算每个单元 i 与其他所有受感染单元 i 的空間连接强度。

Ω_k 是单元格 K 的邻域被感染程度:

$$\Omega_k = w \sum_{i=1}^m \text{con}(S_i = \text{Urbannized})$$

其中, m 表示领域内总的单元数量, S_i 表示相单元 i 的状态,即如果该像元被感染,则 $\text{con}(S_i = \text{Urbannized})$ 返回 1。否则,它的值为 0。由于小班的不规则和离散分布,分别统计了与该小班直接相邻、1km 范围内和 2km 范围内被感染的小班数量,通过 Ω_k 来评估邻域单元格内被感染程度。 w 作为一个可调节参数,以指数的计算形式来突出不同空間范围内邻域状态的对于松材线虫并发生的作用。

表 2 元胞自动机和其他机器学习模型

Tab. 2 Cellular automata and other machine learning models

预测模型 Predict Model	模型简称 Model Name
元胞自动机模型 Cellular automata	CA
支持向量机模型 Support vector machines	SVM
随机森林模型 Random Forest	RF
分类与回归树 Classification and regression tree	CART
广义线性模型 Generalize linear model	GLM
人工神经网络模型 Artificial Neural Network	ANN

为了展示模型预测结果的有效性,实验过程中同时采用元胞自动机模型和其他五种(表 2)主流的机器学习算法进行预测,对结果结果比较。输入的训练数据包含 2021 和 2022 年

松材线虫疫情发生的状态数据以及 1.1.2 节所采集的所有影响因子数据。2023 年数据用与模型验证。

2.3 灰狼优化算法

最后，在训练过程中，为了确定元胞自动机中参数的最优配置，使用灰狼优化算法。灰狼优化算法 (Grey Wolf Optimizer, GWO) 作为一种启发式优化技术，由 Mirjalili 等人于 2014 年提出^[28]。通过模仿灰狼的社会等级和狩猎策略来寻找问题的最优解，在很多建模工作中常常采用灰狼优化算法对模型参数进行估计^[29, 30]，相比于其他优化算法具有更快的收敛速度与求解精度。在历史数据拟合过程中，可以通过计算混淆矩阵来定义适应度函数，利用灰狼优化算法来优化元胞自动机模型中的参数，以达到更高的预测精度和更稳定的系统行为。

GWO 算发数学模型基于以下公式，模拟灰狼的社会行为和狩猎策略：

$$\vec{D} = |\vec{C} \cdot \vec{X}_{prey} - \vec{X}|$$

$$\vec{X}(t+1) = \vec{X}_{prey} - \vec{A} \cdot \vec{D}$$

其中， $\vec{X}$ 是灰狼的位置， $\vec{X}_{prey}$ 是猎物的位置。向量 $\vec{A}$ 和 $\vec{C}$ 是系数向量，由以下公式计算：

$$\vec{A} = 2 \cdot \vec{a} \cdot \vec{r}_1 - \vec{a}$$

$$\vec{C} = 2 \cdot \vec{r}_2$$

这里， $\vec{a}$ 在迭代过程中线性减少从 2 到 0， $\vec{r}_1$ 和 $\vec{r}_2$ 是在 $[0, 1]$ 范围内生成的随机向量。

通过调整参数和迭代，GWO 可以有效地寻找全局最优解。这种算法特别适用于具有复杂、非线性和多峰的优化问题，已被广泛应用于工程、经济和科学研究中。

2.5 模型评价

分类任务的模型预测结果主要可以分为四类：TN、TP、TP 和 FN。归纳为混淆矩阵。准确率 (Accuracy)：预测正确的样本数占总样本数的比例。精确率 (Precision)：预测为正类的样本中实际为正类的比例。召回率 (Recall)：实际为正类的样本中被预测为正类的比例。F1 分数 (F1 Score)：精确率和召回率的调和平均数，提供两者之间的平衡。

使用 6 种模型对 2021 和 2022 年的数据进行拟合，计算混淆矩阵来展示模型在 2023 年验证集上的预测效果 (图 6)，并计算准确率 (Accuracy)、精确率 (Precision)、召回率 (Recall) 和 F1 分数值等性能指标评估模型的性能 (表 3)。

		Predict class 预测类别		
		Positive 正类	Negative 负类	Recall 召回率
True class 真实标签	Positive 正类	229	56	0.804
	Negative 负类	498	25654	0.981
	Precision 准确率	0.315	0.998	0.453

(a) CA 模型

		Predict class 预测类别		
		Positive 正类	Negative 负类	Recall 召回率
True class 真实标签	Positive 正类	42	243	0.147
	Negative 负类	636	25516	0.976
	Precision 准确率	0.062	0.991	0.088

(b) SVM 模型

		Predict class 预测类别		
		Positive 正类	Negative 负类	Recall 召回率
True class 真实标签	Positive 正类	77	208	0.270
	Negative 负类	72	26080	0.997
	Precision 准确率	0.517	0.992	0.355

(c) RF 模型

		Predict class 预测类别		
		Positive 正类	Negative 负类	Recall 召回率
True class 真实标签	Positive 正类	85	200	0.298
	Negative 负类	151	26001	0.994
	Precision 准确率	0.360	0.992	0.326

(d) CART 模型

		Predict class 预测类别		
		Positive 正类	Negative 负类	Recall 召回率
True class 真实标签	Positive 正类	78	207	0.273
	Negative 负类	28	26124	0.999
	Precision 准确率	0.736	0.992	0.398

(e) GLM 模型

		Predict class 预测类别		
		Positive 正类	Negative 负类	Recall 召回率
True class 真实标签	Positive 正类	82	203	0.288
	Negative 负类	35	26117	0.999
	Precision 准确率	0.701	0.992	0.408

(f) ANN 模型

图 6 模型在验证集的分类结果

Fig. 6 Classification results of models in the validation dataset

分析以上 6 种模型在验证集上的分类结果可以发现，F1 分数的排名为 CA>ANN>GLM>RF>CART>SVM，F1 分数综合了模型的准确率和召回率，可以很好地显示出模型的整体性能。CA 模型在召回率方面明显要优于其他 5 种机器学习模型，这说明通过有效的空间信息加入和参数设计，模型对于新发样本的预测性能有了显著提升。而对于其他 5 种机器学习模型来说，不能有效地对正样本即新发小班进行预测。

表 3 模型的性能指标

Tab. 3 Performance indicators of models

模型简称 Model Name	准确率 Accuracy	精确率 Precision	召回率 Recall	F1 分数 F1 Score
CA	0.979	0.315	0.804	0.453
SVM	0.967	0.062	0.147	0.088
RF	0.989	0.517	0.270	0.355
CART	0.987	0.360	0.298	0.326
GLM	0.991	0.736	0.273	0.398
ANN	0.991	0.701	0.288	0.408

3 讨论

基于第二章中所提出的灰狼优化算法—元胞自动机模型,使用训练过程中的最佳参数组合,对2024年江苏省的松材线虫病小班发生情况进行预测,结果在Arcgis中展示(图7)。

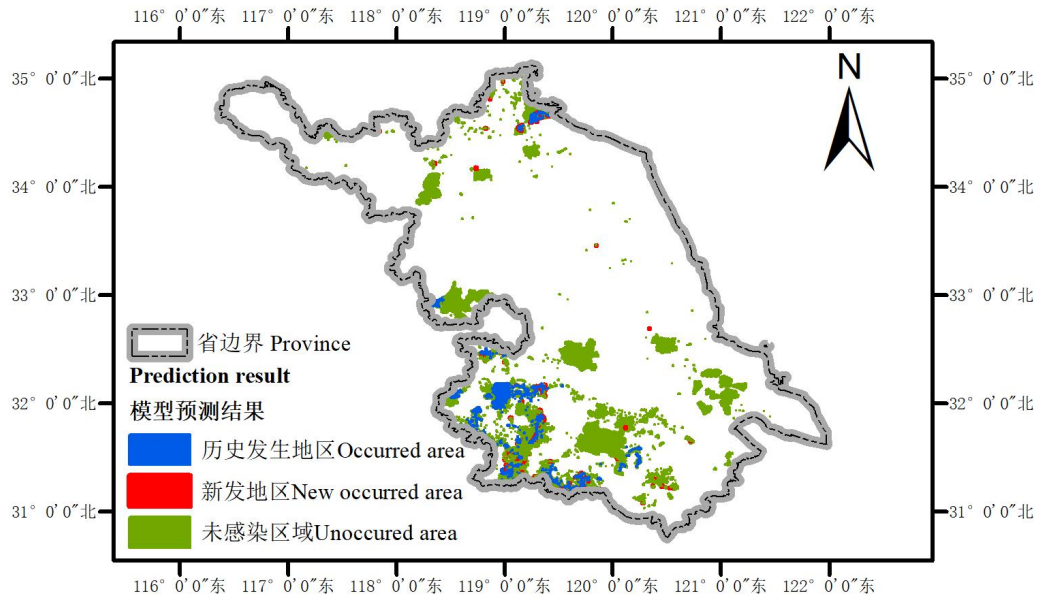


图 7 2024 年江苏省新发小班预测

Figure 7 Predict the occurrence of plaque in Jiangsu Province in 2024

根据预测结果的可视化展示,松材线虫的大部分新发地区主要集中在旧疫情发生地区周围,呈现较强的空间相关性和时空聚集特征^[31]。除此以外,部分地区出现跳跃式传播的情况,这种现象的发生除了直接受到木材调运的影响,还与人口、GDP、道路密度等间接反映人类活动强度的因素有关,人为影响因素在松材线虫的远距离扩散中起着重要作用,这与文献^[32]的观点是一致的,应当采取控制措施来阻止或减缓其扩张。整体来看,疫情的发生,集中在江苏省东南部和西北部地区,呈现横向扩散传播趋势。在短时间内,江苏省中部地区都处于较低的发生风险范围内。通过对疫情新发地区进行预测,可以为之后的疫情防控提供侧重方向,为制定治理措施提供理论支持,减少生态和经济上的损失。

本研究使用了贝叶斯估计方法,探讨了气温、降水、海拔、人口、GDP、道路密度、木材调运这些影响因素对于松材线虫病发生的特征重要性,同时提出了一种入侵物种传播预测的新方法,并与其他主流的机器学习模型进行了比较研究。结果表明,元胞自动机模型的召回率达到了80.4%,预测结果的精确率为31.5%,模型的综合表现优于其他所有机器学习模型。与其他机器学习模型相比,有效地对空间信息进行了利用,在模型中纳入了与其他地区之间的空间连接关系,这种空间特征信息的引入,有效地提升了模型对于正样本的预测效果。

本研究所提出的元胞自动机方法,在新发地区预测中展示出了良好的性能,但是在模型预测的准确性方面还存在问题,主要表现为模型进行了过量的预测,出现这种问题的原因有可能是因为训练样本时间跨度过短而导致的。在以后的工作中,应当更侧重于提升模型预测效果的准确性,通过收集更长时间序列的数据或者继续深入挖掘各影响因子与发生情况之间的关系,有效对发生和不发生地区进行分类。最后值得指出的是,本研究使用了江苏省为例,使用元胞自动机模型对松材线虫病在小班尺度下的数据进行了预测,该方法同样可以用于其他省份未来松材线虫病高风险发生地区的预测。

4 结论

本文所提出的基于灰狼优化算法—元胞自动机模型在模拟中国江苏生松材线虫病的历史传播扩散趋势中取得了较好的结果,为入侵物种的传播预测提供了新的思路。元胞自动机模型作为一种结合空间相互作用和时间关系的动力学模型,能够针对入侵物种的扩散过程进行建模,进行同时结合灰狼优化方法,成为一种结构开放、易于优化和设计的时空动力学模型,应用于复杂的小班尺度数据集中,相比与其他 5 种主流的机器学习模型表现出了更好的性能,能够准确地识别和预测高风险发生小班。本文所提出的方法不仅为入侵物种的预测提供了思路和方法,同时为未来林业工作者们制定防控措施提供了理论依据和技术支持,这对于中国松材线虫病以及其他入侵物种的管理具有重要意义。

参 考 文 献

- [1] 叶建仁.松材线虫病在中国的流行现状、防治技术与对策分析[J].林业科学,2019,55(09):1-10.
- [2] 陈凤毛,汤坚,叶建仁.松材线虫病鉴定方法与评价(综述)[J].安徽农业大学学报,2005,(01):22-25.
- [3] Ju Y W, Li M Y, Wu W H. Predictive methods of pine wilt disease in Jiangsu Province[J]. Sci. Silvae. Sin, 2010, 46(12): 91-96.
- [4] Lee D S, Choi W I, Nam Y, et al. Predicting potential occurrence of pine wilt disease based on environmental factors in South Korea using machine learning algorithms[J]. Ecological informatics, 2021, 64: 101378.
- [5] Mohd Pozi M S, Sulaiman M N, Mustapha N, et al. A new classification model for a class imbalanced data set using genetic programming and support vector machines: case study for wilt disease classification[J]. Remote Sensing Letters, 2015, 6(7): 568-577.
- [6] Zhang Y, Dian Y, Zhou J, et al. Characterizing spatial patterns of pine wood nematode outbreaks in subtropical zone in China[J]. Remote Sensing, 2021, 13(22): 4682.
- [7] 韩阳阳,王焱,项杨,等.基于 Maxent 生态位模型的松材线虫在中国的适生区预测分析[J].南京林业大学学报(自然科学版),2015,39(01):6-10.
- [8] 刘晓梅,蒲永兰,李宏群,等.基于 MaxEnt 模型的重庆松材线虫病潜在生境分析[J].三峡生态环境监测,2017,2(03):75-80.
- [9] Yanlong G, Zefang Z, Huijie Q, et al. Challenges and development trend of species distribution model[J]. Advances in Earth Science, 2020, 35(12): 1292.
- [10] Tang X, Yuan Y, Li X, et al. Maximum entropy modeling to predict the impact of climate change on pine wilt disease in China[J]. Frontiers in plant science, 2021, 12: 652500.
- [11] Liu C, Newell G, White M. On the selection of thresholds for predicting species occurrence with presence-only data[J]. Ecology and evolution, 2016, 6(1): 337-348.
- [12] Zhou H W, Xie M, Koski T M, et al. Epidemiological model including spatial connection features improves prediction of the spread of pine wilt disease[J]. Ecological Indicators, 2024, 163: 112103.
- [13] Aburas M M, Ho Y M, Ramli M F, et al. The simulation and prediction of spatio-temporal urban growth trends using cellular automata models: A review[J]. International Journal of Applied Earth Observation and Geoinformation, 2016, 52: 380-389.
- [14] Medrek M, Pastuszak Z. Numerical simulation of the novel coronavirus spreading[J]. Expert systems with applications, 2021, 166: 114109.
- [15] Liu D, Zhang X. Occurrence Prediction of Pine Wilt Disease Based on CA-Markov Model[J]. Forests, 2022, 13(10): 1736.
- [16] Liu Y, Huang J, Yang T. Natural Factors Play a Dominant Role in the Short-Distance Transmission of Pine Wilt Disease[J]. Forests, 2023, 14(5): 1059. 许格希,余荣兵,杨昌旭,等.基于 GIS 空间技术和 MaxEnt 模型预测川西松材线虫病入侵风险[J].北京林业大学学报,2023,45(09):104-115.
- [17] Wang W, Zhu Q, He G, et al. Impacts of climate change on pine wilt disease outbreaks and associated carbon stock losses[J]. Agricultural and Forest Meteorology, 2023, 334: 109426.
- [18] Xiang YiJuan X Y J, Guan JianFeng G J F, Li Jian L J, et al. A pine-wilt disease forecasting model for Zhejiang Province[J]. 2011.

- [19] Aburas M M, Ho Y M, Ramli M F, et al. The simulation and prediction of spatio-temporal urban growth trends using cellular automata models: A review[J]. *International Journal of Applied Earth Observation and Geoinformation*, 2016, 52: 380-389.
- [20] Wang W, Zhu Q, He G, et al. Impacts of climate change on pine wilt disease outbreaks and associated carbon stock losses[J]. *Agricultural and Forest Meteorology*, 2023, 334: 109426.
- [21] Azarafza M, Azarafza M, Akgün H, et al. Deep learning-based landslide susceptibility mapping[J]. *Scientific reports*, 2021, 11(1): 24112.
- [22] 李明阳, 张晓利, 刘方, 等. 基于贝叶斯最大熵模型的紫金山松材线虫危害程度时空分析[J]. *西北农林科技大学学报(自然科学版)*, 2012, 40(07): 99-105.
- [23] Lee D S, Choi W I, Nam Y, et al. Predicting potential occurrence of pine wilt disease based on environmental factors in South Korea using machine learning algorithms[J]. *Ecological informatics*, 2021, 64: 101378.
- [24] Beroho M, Briak H, Cherif E K, et al. Future scenarios of land use/land cover (LULC) based on a CA-markov simulation model: case of a mediterranean watershed in Morocco[J]. *Remote Sensing*, 2023, 15(4): 1162.
- [25] Lukas P, Melesse A M, Kenea T T. Prediction of future land use/land cover changes using a coupled CA-ANN model in the upper omo-gibe river basin, Ethiopia[J]. *Remote Sensing*, 2023, 15(4): 1148.
- [26] Lin J, Li X, Wen Y, et al. Modeling urban land-use changes using a landscape-driven patch-based cellular automaton (LP-CA)[J]. *Cities*, 2023, 132: 103906.
- [27] Yu B, Yin H, Zhu Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting[J]. *arXiv preprint arXiv:1709.04875*, 2017.
- [28] Mirjalili S, Mirjalili S M, Lewis A. Grey wolf optimizer[J]. *Advances in engineering software*, 2014, 69: 46-61.
- [29] Jin Z, He D, Wei Z. Intelligent fault diagnosis of train axle box bearing based on parameter optimization VMD and improved DBN[J]. *Engineering Applications of Artificial Intelligence*, 2022, 110: 104713.
- [30] Zhou J, Qiu Y, Zhu S, et al. Optimization of support vector machine through the use of metaheuristic algorithms in forecasting TBM advance rate[J]. *Engineering Applications of Artificial Intelligence*, 2021, 97: 104015.
- [31] Hao Z, Huang J, Li X, et al. A multi-point aggregation trend of the outbreak of pine wilt disease in China over the past 20 years[J]. *Forest Ecology and Management*, 2022, 505: 119890.
- [32] Robinet C, Roques A, Pan H, et al. Role of human-mediated dispersal in the spread of the pinewood nematode in China[J]. *PLoS One*, 2009, 4(2): e4646.