

引用格式:李照,高惠瑛,代晓奕,等.一种耦合LSTM算法和云模型的疫情传播风险预测模型[J].地球信息科学学报,2021,23(11):1924-1935.
[Li Z, Gao H Y, Dai X Y, et al. An epidemic spread risk prediction model coupled with LSTM algorithm and cloud model[J]. Journal of Geo-information Science, 2021,23(11):1924-1935.] DOI:10.12082/dqxxkx.2021.210576

一种耦合LSTM算法和云模型的疫情传播风险预测模型

李 照,高惠瑛,代晓奕,孙 海

中国海洋大学工程学院 城市与工程管理信息化山东高校重点实验室,青岛 266100

An Epidemic Spread Risk Prediction Model Coupled with LSTM Algorithm and Cloud Model

LI Zhao, GAO Huiying, DAI Xiaoyi, SUN Hai

Key Laboratory of Modern Management Informationization Universities in Shandong, College of Engineering, Ocean University of China, Qingdao 266100, China

Abstract: The COVID-19 epidemic poses a great threat to public health and people's lives, which has initiated new challenges to the prevention and control system of the epidemic in China. In all efforts for epidemic control and prevention, predicting the risk of epidemic spread is of great practical importance for scientific prevention and control, and precise strategies. To predict the risk of an epidemic rapidly and quantitatively, this paper fused multi-source spatiotemporal data and established a risk prediction model for epidemic transmission by coupling LSTM algorithm and cloud model. Firstly, a simulation model of the spatiotemporal spread of infectious diseases was built based on GIS and LSTM algorithm, which simulated the infectious disease's spatiotemporal transmission process by learning rules in historical epidemic data. At the same time, to improve the simulation accuracy, this paper took $1\text{ km} \times 1\text{ km}$ for the spatial scale, and days for the temporal scale as the study scale. Secondly, this paper applied the simulated data of infectious cases and the spatiotemporal influence factors on the spread of the epidemic to construct risk evaluation indicators. Finally, the cloud model and adaptive strategies were applied to construct an epidemic risk assessment model. In this way, the epidemic risk assessment at multiple spatial scales was achieved. In the empirical study phase, based on the Beijing COVID-19 epidemic data from 11 June 2020 to 25 June 2020, this paper simulated the process of the spatial evolution of the epidemic from 26 June 2020 to 1 July 2020. To test the advantage of the LSTM model applied to simulate spatiotemporal spread of infectious diseases, four machine learning models were introduced for comparison, including GA-BP Neural Network, Decision Regression Tree, Random Forest, and Support Vector Machine. The results were as follows: ① Compared with other conventional machine learning models, the LSTM model with time-series relationship had higher simulation accuracy ($MAE=0.002\ 61$) and better fitting degree ($R\text{-Square}=0.9455$). This

收稿日期 2021-09-25;修回日期:2021-11-09.

基金项目 国家自然科学基金项目(41906185、U1706226、52071307)。[**Foundation items:** National Natural Science Foundation of China, No.41906185, No.U1706226, No.52071307.]

作者简介 李 照(1996—),女,山东济南人,硕士生,主要从事城市突发公共卫生防灾及风险评估研究。

E-mail: zz2015@stu.ouc.edu.cn

*通讯作者 高惠瑛(1967—),女,山东青岛人,博士,教授,主要从事城市灾害的风险管理以及城市安全管理信息系统研发。

E-mail: fqmgghy@sina.com

showed that the LSTM model considering the temporal relationship between epidemic data was more suitable for epidemic spatial evolution simulation. ② The application results showed that the coupled model can not only fully consider the influence of infection source factors, weather factors, epidemic spread factors and epidemic prevention factors on the spread of transmission risk and reflect the trend of risk evolution, but also quickly quantify regional risk levels. Therefore, the coupled model based on LSTM algorithm and cloud model can effectively predict the transmission risk of epidemic, and also provide a method reference for establishing spatial-temporal transmission models and assessing epidemic risk.

Key words: infectious diseases; Long Short-Term Memory (LSTM) model; cloud model; spatial evolution simulation; risk assessment; risk prediction; coupled model; COVID-19

***Corresponding author:** GAO Huiying, E-mail: fqmghy@sina.com

摘要 模拟传染病时空传播、定量评估疫情风险对科学防控、精准施策具有重要的现实意义。本文融合多源时空数据,构建了耦合 LSTM 算法和云模型的疫情传播风险预测模型。该模型首先基于 GIS 和 LSTM 算法构建疫情空间演变模拟模型,通过学习历史疫情数据中的规律,以 $1\text{ km}\times 1\text{ km}$ 为空间尺度、天为时间尺度模拟传染病时空传播过程。其次,基于模拟传染病例数据和疫情传播时空影响因素构建风险评价指标,应用云模型和自适应策略构建疫情风险评估模型,实现多空间尺度的疫情风险评估。在实证研究阶段,应用该模型对北京 2020 年 6 月份突发 COVID-19 疫情空间演变过程进行模拟和风险评估,并引入常规机器学习模型作比较验证。结果表明:应用于疫情时空传播模拟,相较其它常规的机器学习模型,考虑时序关系的 LSTM 模型的模拟精度更高(MAE 为 0.00261),拟合度更好($R\text{-square}$ 为 0.9455);耦合模型不仅能充分考虑传染源因素、天气因素、疫情扩散因素及疫情防御因素对疫情风险传播的影响,反映风险演变趋势,还能快速量化区域风险等级,实现不同空间分辨率下的疫情风险评估。因此,基于 LSTM 算法和云模型的耦合模型可有效预测疫情的传播风险,同时,也为传染病时空传播建模与风险评估提供了方法参考。

关键词 传染病;长短期记忆(LSTM)模型;云模型;空间演变模拟;风险评估;风险预测;耦合模型;COVID-19

1 引言

新冠肺炎疫情是近代以来感染规模最大、影响最为深远的一次突发公共卫生事件,其造成的灾难性后果、暴露出疫情防控中的弊端以及存在的潜在威胁,向我国的疫情防控体系发起了新的挑战。在疫情防控的各项工作中,模拟病毒传播、快速评估疫区风险是科学判断疫情、精准把控疫情的重要依据。因此,有必要在疫情发生时应用科学的方法识别潜在疫区、精细化评估疫情风险。

如今,随着信息技术的发展,具备空间位置属性的生态要素数据和社会要素数据增长迅速,汇集丰富的时空数据集,使构建科学有效的疫情传播风险预测模型成为可能。基于多源时空数据预测疫情传播风险,我们首先需要模拟精细空间尺度下的疫情空间演变过程。目前学者主要通过建立动力学模型和统计学模型实现对传染病时空传播的模拟。其中,仓室模型在动力学模型中一直占据着主流地位,是疫情模拟和预测研究的重要方法之一^[1-2],仓室模型可分为 SI^[3]、SIS^[4]、SIR^[5]、SEIR^[6-8]等基本类型。仓室模型侧重于研究传染病传播的机理,从而

反映出疫情传播的内在规律,可以考虑人口流动、空间异质性以及结合具体场所数据,修改模型参数及设定,在微观尺度上模拟传染病时空扩散过程^[9-11],但其建模复杂度和对数据粒度的细化标准相对较高,确诊病例的空间分布模拟还具有较大的不确定性。统计学模型侧重于学习历史数据中的规律,从而实现对疫情传播趋势的模拟和预测,相比传统的动力学模型,统计学模型参数相对较少,应用更加简便高效,如经验贝叶斯模型^[12]、自回归综合移动平均(ARIMA)模型^[13-14]以及 Logistic 增长曲线模型^[15]在传染病流行时间以及规模研究中的预测效果较好。此外,随着近几年人工智能的发展,将历史数据作为训练数据集,应用机器学习模拟疫情传播的预测方法也得到了快速发展,已广泛应用于疫情发展趋势及停止时间的模拟预测。研究表明,机器学习方法可以更好地模拟疫情传播与时空影响因素的关系^[16-17],相比 SEIR 等动力学模型,建模时需要参数(如基本再生数 R_0)进行假设而导致模型模拟精度低的问题,只需借助实时数据就能计算模型参数的机器学习方法在非线性系统演变模拟中展现出巨大的优势^[18],结合地理信息系

统的机器学习演变模型打开了传染病疫情空间演变模拟研究的新思路。如BP神经网络^[19-21]、支持向量机模型^[22-23]、随机森林模型^[24-25]已初步尝试应用于疫情时空传播模拟,结果显示真实值和模拟值间的误差较小,表明了机器学习方法在传染病时空扩散模拟应用中的可行性。

虽然以上机器学习方法解决了传统数学模型在非线性系统建模的局限性,但在建模时,输入变量常被认为是彼此独立的,忽略了疫情数据是一个时间序列数据集,数据间存在时序关系,疫情空间演变存在随趋势变化的情况,所以应用于疫情传播风险预测,模型模拟的精度还有待提高。此外,虽然目前针对传染病风险评估的研究已取得不错的成果,但是仍存在只将确诊病例作为疫情风险评估的单一评价指标,风险评估结果不够客观、准确,无法反映风险演变趋势的不足^[8],基于模拟结果和疫情传播时空影响因素,如何量化的描述评估单元的风险等级,快速、客观地为疫情防控工作提供决策支持的研究还较少。

鉴于此,本文提出了耦合长短期记忆(Long Short-term Memory, LSTM)算法和云模型的疫情传播风险预测模型。相较已有研究,本文所提模型有如下改进:①对于传染病确诊病例点要素,考虑距离衰减效应的影响,应用核密度分析处理模拟模型

的输入数据,并以1 km×1 km为空间尺度、以天为时间尺度来研究,提高模拟精细程度;②将深度学习算法——LSTM算法引入疫情时空建模的构建中来,并与常规机器学习模型作比较,验证LSTM模型模拟疫情时空演变的优势;③在模拟结果的基础上展开更深层的研究,构建LSTM算法和云模型的耦合模型,基于模拟确诊病例数据和疫情传播时空影响因素构建风险评价指标,应用自适应策略,实现了不同空间分辨率下(县(区)、乡镇(街道)、1 km格网尺度)的疫情风险定量评估,从而预测出各区域疫情传播的风险。

2 研究方法

本文提出模型的设计思路为:①数据层:通过网站设置的API接口以及Python爬虫等渠道获取相关影响因素数据;②分析处理层:借助GIS,将数据层的数据,以1 km×1 km格网为单位进行分析、处理,形成模型层的输入数据;③模型模拟层:将处理后的输入数据输入LSTM神经网络模型,训练模拟传染病疫情的空间演变过程;④风险评估层:输入疫情模拟结果与疫情风险评价指标,应用云模型和自适应策略,输出疫情风险评价结果。模型构建流程如图1所示,模型的主要构建过程简述如下。

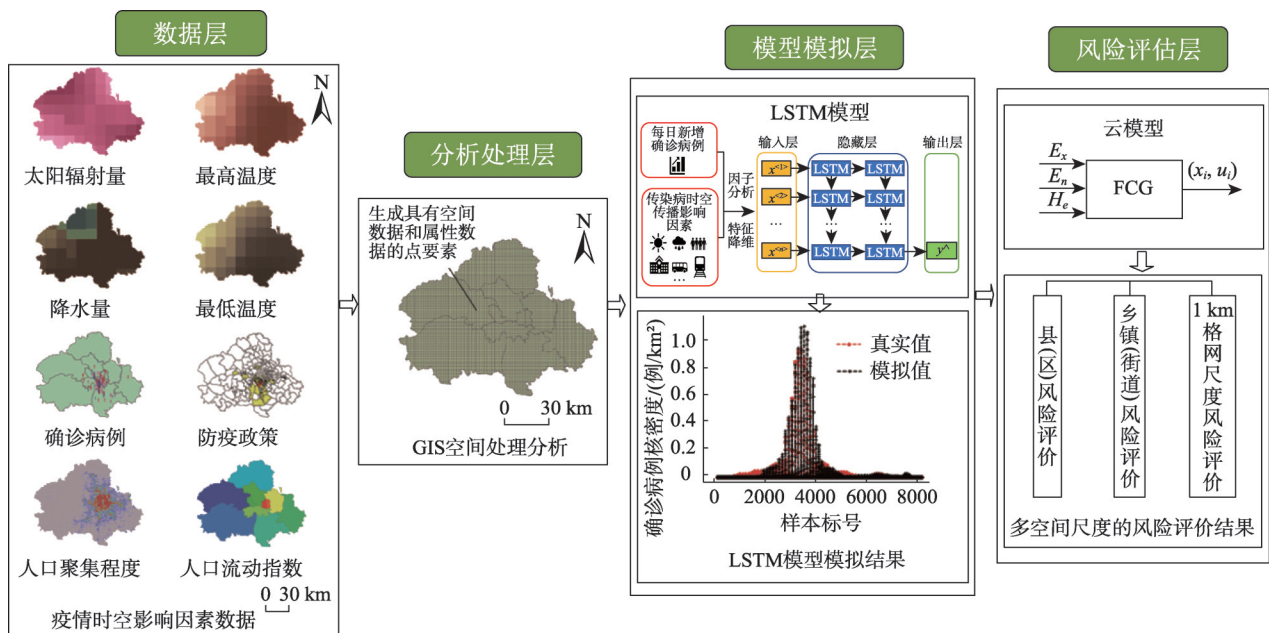


图1 模拟流程图

Fig. 1 Simulation flow chart

2.1 基于 GIS 的数据分析处理

地理信息系统技术凭借着强大的制图功能和空间分析、统计功能成为国内外学者研究疫情时空演变规律的重要基础工具^[26],被广泛应用于埃立克体病、非典型性肺炎(SARS)、疟疾、H5N1 高致病性禽流感、登革热、新冠肺炎疫情等流行传染病的时空演变规律的研究。因此,本文基于数据层得到影响传染病传播的相关影响因素数据后,借助 GIS 对数据进行分析处理。主要方法有:

(1)应用空间可视化功能,将影响疫情演变的时空数据可视化呈现。同时,为了更精确地模拟传染病的传播过程,将疫区划分为 $1\text{ km} \times 1\text{ km}$ 的网格,以此为单位进行研究。

(2)对各影响因素量化处理。其中确诊病例点要素,应用核密度分析工具分析处理,核密度分析方法不同于常见的密度分析方法,它在计算某一点 S 处的核密度值时考虑了距离衰减效应的影响^[27]。点 S 处的核密度值可表示如下:

$$\lambda(s) = \sum_{i=1}^n \frac{1}{\pi r^2} k\left(\frac{d_{is}}{r}\right) \quad (1)$$

式中: $\lambda(s)$ 表示点 S 处的核密度值; r 表示搜索半径; d_{is} 表示点 i 到点 S 的距离; $k()$ 表示核函数。计算得出的核密度值越大,表示一定距离内所含的样本点数量越多,即确诊病例点数量越多。相比传统的将点数量作为指标分析时,只能反映该区域的情况,却不能反映对周围区域的影响,应用核密度分析既能反映该区域确诊病例点的密集程度,也能反映该区域与确诊病例点的距离远近,更能准确表达各因素在传染病传播中产生的作用。

(3)应用叠加分析功能,形成空间样本点,使每个样本点同时具有空间数据和各影响因素的属性数据(如人口聚集程度、温度等)。但由于输入特征的信息存在重叠时会对模型训练的收敛速度和精度产生影响,因此本文应用因子分析提取共性特征作为模型的输入特征,形成模型层的输入数据。

2.2 基于 LSTM 模型的传染病扩散模拟

近年来,深度学习技术逐渐成熟,凭借着出色的非线性映射能力和学习能力,被广泛应用于众多领域。其中,作为递归神经网络(Recursive Neural Network, RNN)一个变种的 LSTM 模型,不仅能够学习历史时序数据中的时间信息进行模拟,还解决了 RNN 模型梯度消失和梯度爆炸的不足,提高其

长期记忆能力。在语音识别、交通流速预测、机器翻译、故障时间序列预测等领域表现出强大的适应能力^[28]。因此,本文将 LSTM 模型引入到疫情空间演变模拟模型的构建中来。

LSTM 模型是由 Hochreiter 等^[29]提出,是循环神经网络的改良版,具有记忆功能,它可以学习历史数据中存在的时间依赖关系,将时间序列上的信息关联起来,从而模拟疫情的空间演变过程,输出模拟时间段新增确诊病例核密度的模拟值。传染病发病率受到多方因素的影响,为了训练 LSTM 模型通过学习历史数据模拟疫情传播与时空影响因素的关系,构建多变量时间序列作为模型输入特征。模型输入特征由每个时间段新增确诊病例数据和量化的传染病时空传播影响因素组成。但考虑到当各因素之间存在相关性以及输入特征过多时,会对模型的学习效果产生不利影响,因此,首先应用因子分析在保留原有数据信息量的同时对原有特征进行降维,将降维后的共性特征作为输入变量。接下来,应用历史疫情数据训练神经网络各层节点之间的连接权值,神经网络各层节点之间的连接权值可以反映各影响因素对传染病时空传播的影响,其连接权值基于模型的最佳学习效果确定。最后,多变量 LSTM 模型通过学习各时空影响因素对新增确诊病例的影响以及时间序列随趋势变化的关系,从而模拟或预测未来的新增确诊病例情况。模型的网络结构如图 2 所示。其中, n 表示模型的输入时间序列长度, $x^{<n>}$ 是一个多变量输入特征向量,表示第 n 个时间段影响传染病传播的各影响因素的特征值, y^{\wedge} 表示模拟时间段新增确诊病例核密度。

2.3 基于云模型和自适应策略的疫情风险评估

由 LSTM 模型得到传染病传播的模拟数据后,为了更直观、准确地给予政府防疫部门提供决策支持,需要充分地评价各区域的疫情风险状况。

云模型作为可以实现定性概念与定量概念的不确定转换模型^[30],可以辅助我们依据模拟结果和开源时空数据实现对各区域的疫情风险定量评价。该模型可用期望值 Ex 、熵 En 和超熵 He 表示模型的特征。其中,期望值 Ex 是将定性概念量化最典型的值,熵 En 表示定性概念的模糊程度,超熵 He 可以度量熵 En 的不确定性。计算公式如下:

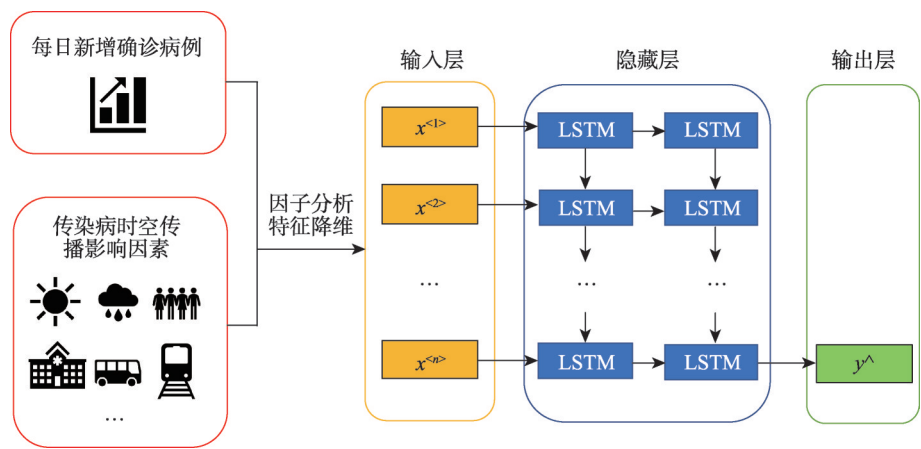


图2 多变量LSTM模型的网络结构
Fig. 2 Network structure of multivariable LSTM model

$$\begin{cases} Ex_i = \frac{z_i^{\min} + z_i^{\max}}{2} \\ En_i = \frac{z_i^{\max} - z_i^{\min}}{3} \\ He = k \end{cases} \quad (2)$$

式中： z_i^{\min} 和 z_i^{\max} 分别是评价指标某一等级标准的最小值与最大值， k 值取经验值。在本文中，根据指标可获得性并参考新冠肺炎疫情传播影响因素的研究^[31-34]，最终选取的评价指标如表1所示，风险评价标准分为五级，分别是：高风险地区、较高风险地

区、中风险地区、较低风险地区以及低风险地区，评价指标权重根据熵权法确定。
此外，为了实现快速评估区域疫情风险，本文采用自适应策略，先评价低空间分辨率下的疫情风险，再针对高风险区以高空间分辨率进一步评价，从而实现多空间尺度的疫情风险评价，满足不同精度的疫情防控要求。

3 实验数据与模型设计

3.1 实验区域及数据来源

本文以北京为研究区域展开疫情传播风险预测模型的实证研究。实验数据来源及类型详见表2。此外，为提高模拟精度，只分析北京市含有确诊病例的区(共11区，图3)。

3.2 实验环境及设计

本研究应用深度学习库 Keras 搭建 LSTM 模型。将2020年6月11日至2020年6月25日共计

表1 疫情风险评价指标

Tab. 1 Epidemic risk assessment index

一级指标	二级指标
传染源因素 A	模拟新增确诊病例核密度 A1
天气因素 B	2 m 处最高温度 B1
	2 m 处最低温度 B2
	总降水量 B3
	总天空直接太阳辐射量 B4
疫情防御因素 C	防疫政策 C1
疫情扩散因素 D	人口聚集程度 D1
	人口流动指数 D2

表2 实验数据

Tab. 2 Experimental data

影响因素	基础数据	来源	数据类型
疫情	新增确诊病例(2020-06-11—2020-07-01,共309条)	北京卫健委	矢量(点)
天气	2 m 处最高温度数据、2 m 处最低温度数据、总降水量数据、总天空直接太阳辐射量数据	欧洲气象中心发布的ERA5资料 ^[35]	0.125°分辨率栅格
人口流动	基于微博签到数据的区人口流动指数 ^[36] (2020-06-11-2020-07-01)	新浪微博发布的位置信息	矢量(面)
人口聚集	百度热力图(2020-06-11—2020-07-01)	百度地图 APP	1 km×1 km 栅格
政策	乡镇(街道)区域风险等级	北京卫健委、北京市疾病预防控制中心	矢量(面)



图3 研究区概况

Fig. 3 Overview of the study area

15 d 的疫情数据作为训练集, 2020 年 6 月 26 日至 2020 年 7 月 1 日共计 6 d 的疫情数据作为测试集。以 3 d 为一个时间段, 用前 3 个时间段的数据作为输入样本, 模拟未来 1 个时间段的新增确诊病例情况。LSTM 模型的参数调试主要包括隐藏层节点个数和模型深度。此外, 新冠肺炎发病率与天气、人口密度、人口聚集、人口流动以及政策等因素有明显的联系^[31-34]。LSTM 模型输入特征具体为: 每个时间段的新增确诊病例核密度、2 m 处最高温度、2 m 处最低温度、总降水量、总天空直接太阳辐射量、人口流动指数、人口聚集程度以及量化的防疫政策(根据区域风险等级进行量化, 低风险区记为“0”, 中风险地区记为“1”, 高风险地区记为“2”)共计 8 个指标因素。实验对比模型分别是 LSTM 模型、GA-BP 神经网络模型、决策树模型、随机森林模型和支持向量机模型。

3.2.1 模型性能评价指标

本研究使用平均绝对误差(MAE)(式(3))来评判模型模拟的偏离程度, 使用决定系数(R-square)(式(4))来评判模型的拟合能力。MAE 越小表示模型模拟精度越高, R-square 越接近 1, 表示模型拟合的效果越好, 输入变量对模拟值的解释能力也越强。

$$MAE = \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i) \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^m (y_i - \hat{y}_i)^2}{\sum_{i=1}^m (y_i - y_{\text{mean}})^2} \quad (4)$$

式中: m 为样本数; y_i 为原始值; \hat{y}_i 为模拟值; y_{mean} 为样本均值。

3.2.2 隐含层节点个数

目前研究中, 隐藏层神经元个数通常情况下按照式(5)选取神经元个数, 其中, N 为隐藏层神经元个数, n 为输入神经元个数, m 为输出神经元个数, a 取 1~10 的常数。

$$N = \sqrt{n + m} + a \quad (5)$$

在本文研究中, 应用因子分析提取可代表所有特征的共性特征数量共计 4 个, 因此, n 为 4, m 为 1, 根据公式 N 应为 3~12。将模型深度设置为 1, 调整隐藏层神经元个数, 根据模型模拟结果的平均绝对误差来衡量模型模拟的偏离程度, 从而确定隐藏层神经元个数。

模拟结果显示, 当 N 调整为 8 时, 模型的 MAE 最小, 模拟效果最好。测试集的模拟结果如表 3 所示。

表3 隐藏层调整的模拟结果

Tab. 3 Simulation results of hidden layer adjustment

隐藏层神经元个数	MAE (6-25—6-28)	MAE (6-29—7-1)	MAE 合计
3	0.002 06	0.001 65	0.003 71
4	0.002 14	0.001 54	0.003 68
5	0.002 12	0.001 69	0.003 81
6	0.002 05	0.001 31	0.003 36
7	0.001 86	0.001 23	0.003 09
8	0.001 85	0.001 02	0.002 87
9	0.001 83	0.001 13	0.002 96
10	0.001 91	0.001 27	0.003 18
11	0.002 12	0.001 22	0.003 34
12	0.002 05	0.001 17	0.003 22

3.2.3 模型深度

鉴于本例研究疫情数据相对较少, 若模型隐藏层数量过高会导致模型过拟合, 因此, 实验设定模型深度为 1~5 层, 每层隐藏神经元数量仍为 8 个, 根据模型模拟结果的平均绝对误差和决定系数来评判模型性能, 从而确定模型深度。根据模拟结果(表 4)可知, LSTM 模型模型深度设置为 2 层时, 模型模拟的偏离程度最小, 拟合的效果最好, 输入影响因素值对模拟值的解释能力也最强。

表4 模型深度调整的模拟结果

Tab. 4 Simulation results of model depth adjustment

层数	MAE	R-square
1	0.00287	0.9391
2	0.00261	0.9455
3	0.00314	0.9361
4	0.00353	0.9206
5	0.00374	0.9132

4 结果及分析

4.1 COVID-19确诊病例模拟空间分布

将 LSTM 模型模拟结果进行空间可视化展示。由疫情的空间分布模拟结果可视化图与真实分布情况对比(图4)可知:疫情呈聚集分布,新增确诊病例主要集中在丰台区,表明丰台区模拟时间段新增确诊病例数量多且聚集程度高;二者空间分布具有较大的相似性,尤其是对新增确诊病例核密度较高的地区识别更为准确,这说明该模型对于疫情聚集的区域模拟精准度较高;新增确诊病例核密度在0~0.1例/km²的空间分布有一定的差异,这一方面是由于模型本身存在一定的误差,另一方面则因为零星确诊病例的分布受多方因素的影响,其空间分布模拟具有较大的不确定性,定量模拟模型难以非常准确的模拟出其空间传播过程。总而言之,该模型对新冠肺炎的空间分布模拟精度较高,能够较为准确地模拟出潜在确诊病例区。

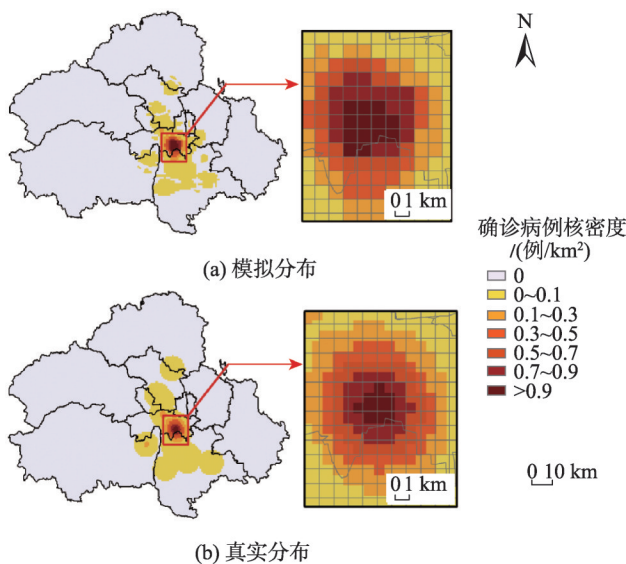


图4 疫情的空间真实分布与模拟分布对比

Fig. 4 Comparison of spatial real distribution and simulated distribution of epidemic situation

4.2 LSTM模型模拟结果及对比验证

经过实验训练, LSTM模型与对比模型在测试集中的模拟结果误差分布如图5所示。为突出本文所提出模型的优势,依据3.2.1节模型性能评价标准,对各模型的模拟效果进行对比,模型性能对比结果如表5所示。

依据模拟结果误差分布图和模型性能对比结果,从模拟结果误差分布来看, LSTM模型测试集中90%的模拟误差值集中分布在区间[-0.02,0.02],

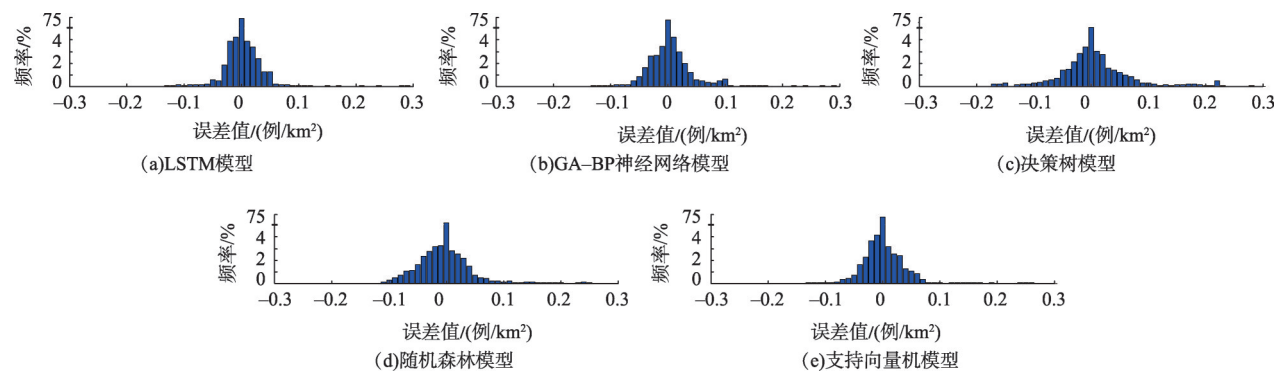


图5 模拟结果误差分布

Fig. 5 Comparison diagram of error distribution

表5 模型模拟结果对比

Tab. 5 Comparison of model simulation results

评判标准	LSTM模型	GA-BP神经网络	决策回归树	随机森林	支持向量机
平均绝对误差	0.002 61	0.003 20	0.008 72	0.007 43	0.003 79
决定系数	0.945 50	0.936 80	0.860 10	0.912 20	0.925 60

GA-BP 神经网络、决策回归树、随机森林以及支持向量机 4 个机器学习模型测试集中 90% 的模拟误差值分别集中分布在区间 $[-0.03, 0.02]$ 、 $[-0.05, 0.04]$ 、 $[-0.04, 0.03]$ 、 $[-0.02, 0.03]$, 相比之下, LSTM 模型的模拟结果与真实结果更为接近, 误差值更小; 从模拟精度来看, LSTM 模型平均绝对误差最小, 精度最高, GA-BP 神经网络模型和支持向量机模型性能次之, 精度最低、误差最大的是随机森林模型和决策树模型; 从拟合效果来看, LSTM 模型能较好的模拟传染病时空传播过程, R -square 最接近 1, 模拟值与真实值之间的拟合效果最好, 其次是 GA-BP 神经网络模型和支持向量机模型, 拟合度较差的是随机森林模型和决策树模型。因此, 在这 5 种机器学习模型中, 考虑数据间时序关系的 LSTM 神经网络模型更适用于疫情的空间分布预测。由此证明 LSTM 模型在考虑多因素的、精细空间尺度下的疫情空间演变模拟中具有一定的可靠性。

4.3 多空间尺度的疫情风险评估

本文所提出的模型可以实现快速、定量地评估多空间尺度的疫情风险, 多空间尺度分 3 级, 分别是: 北京县(区)、乡镇(街道)和 1 km 格网尺度。另外, 由于在 2020 年 6 月 26 日—7 月 1 日政府部门公示的疫情风险等级查询结果只精确至乡镇(街道), 因此本文仅以北京县(区)和乡镇(街道)为评估单元进行模型可靠性的验证。

4.3.1 县(区)风险评价

根据 2.3 节所述公式, 云模型的计算参数矩阵如表 6 所示。对于上文提到的每个指标, 根据表 6 所示的云模型参数矩阵, 我们可以通过前向云生成器生成隶属度矩阵。考虑到计算结果的随机性, 我们计算了 1000 次以获得更高的精度。在得到各指标的隶属度矩阵后, 结合熵权法计算出的权重系数, 可以对北京市含确诊病例的 11 个县级区域进行疫情风险综合评价。疫情风险评估结果如图 6 所

表 6 云模型的计算参数矩阵

Tab. 6 Calculation parameter matrix of cloud model

指标	低风险	较低风险	中风险	较高风险	高风险
A1	(0.0005, 0.0003, 0.1)	(0.0030, 0.0013, 0.1)	(0.0065, 0.0010, 0.1)	(0.0540, 0.0307, 0.1)	(0.1100, 0.0067, 0.1)
B1	(146.42, 97.61, 0.1)	(293.97, 0.76, 0.1)	(296.08, 0.65, 0.1)	(298.14, 0.72, 0.1)	(300.15, 0.62, 0.1)
B2	(144.63, 96.42, 0.01)	(290.35, 0.72, 0.01)	(292.29, 0.57, 0.01)	(294.10, 0.64, 0.01)	(295.83, 0.52, 0.01)
B3	(0.0008, 0.0005, 0.1)	(0.0020, 0.0002, 0.1)	(0.0029, 0.0003, 0.1)	(0.0041, 0.0004, 0.1)	(0.005 9, 0.000 8, 0.1)
B4	(1 288 026, 858 684, 0.1)	(2 691 607, 77037, 0.1)	(2 902 592, 63620, 0.1)	(3120 919, 81 931, 0.1)	(3 379 396, 90 387, 0.1)
C1	(0.1500, 0.1000, 0.01)	(0.4000, 0.0667, 0.01)	(0.6000, 0.0667, 0.01)	(0.7500, 0.0333, 0.01)	(0.850 0, 0.0333, 0.01)
D1	(0.5000, 0.3333, 0.01)	(1.5000, 0.3333, 0.01)	(2.5000, 0.3333, 0.01)	(4.0000, 0.6667, 0.01)	(6.000 0, 0.6667, 0.01)
D2	(0.4409, 0.2940, 0.01)	(1.5724, 0.4603, 0.01)	(3.1944, 0.6210, 0.01)	(5.3173, 0.7942, 0.01)	(8.2675, 1.1726, 0.01)

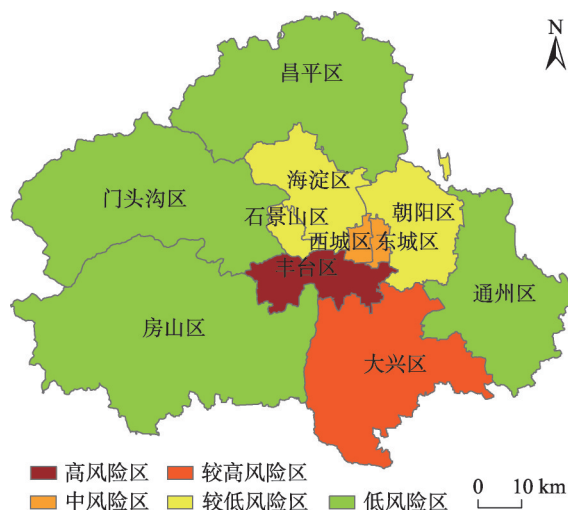


图 6 北京市含确诊病例县(区)风险评价结果

Fig. 6 Risk assessment results of counties (districts) with confirmed cases in Beijing

示。丰台区的风险最高, 其次, 风险较高的是大兴区, 中风险区包括西城区和东城区 2 个区, 较低风险区包括石景山区、海淀区和朝阳区, 其余区则属于低风险区。这与北京的实际疫情风险状况基本相符, 据北京市新冠肺炎疫情防控工作新闻发布会对高中风险地区变化的回复以及国务院客户端小程序的“疫情风险等级查询”结果 (<https://mp.weixin.qq.com/s/0MH-z-hWgvNrXjGqgRF5BQ>, <https://m.weibo.cn/1893892941/4519751785797622>, 于 2021 年 6 月 30 日访问), 在 2020 年 6 月 26 日—7 月 1 日期间, 丰台区有 2 个街道一直处于高风险区, 大部分街道一直处于中风险区; 大兴区有 2 个街道一直处于高风险区, 较少部分街道一直处于中风险区; 西城区、东城区、海淀区有部分街道长时间处于中风险区; 朝阳区和石景山区较少部分街道长时间处于中

风险区;其余区则近乎所有区域处于低风险区。综上所述,此次疫情风险主要集中在丰台区及周边区域,实际情况与我们的评价结果吻合较好。

4.3.2 乡镇(街道)风险评价

经过对北京区域疫情风险的初步评估,我们得知丰台区是疫情传播的高风险区,由此,我们可以根据自适应策略,针对丰台区以乡镇(街道)为评估单元作更详细的风险评估,而不需对每一个区域以高空间分辨率进行风险评价,从而大大节省了时间成本,有利于快速给予政府防疫工作部门决策支持。由疫情风险评估可视化图(图7)可知,花乡、卢沟桥以及丰台街道是疫情防控重点地区,需加强疫情防控政策,减少疫情扩散风险,此外,太平桥街

道、马家堡街道以及南苑乡虽然疫情风险性较小,但仍不可放松警惕,要防范疫情二次传播的风险。在模拟阶段,实际乡镇(街道)疫情风险等级的情况为:花乡一直处于高风险区;卢沟桥、丰台街道、南苑、马家堡街道、右安门街道一直处于中风险区;西罗园街道、太平桥街道、大红门街道、长辛店镇则随着疫情防控措施的落实,发病趋势逐渐平稳并回落,变为低风险区;其余区一直为低风险区。疫情风险评估结果与实际情况基本相符。

4.3.3 1 km 格网尺度风险评价

以1 km 格网为空间尺度评估北京含确诊病例区的疫情风险,得到的风险评价结果如图8所示。其中,高风险区主要集中分布在丰台区,较高风险

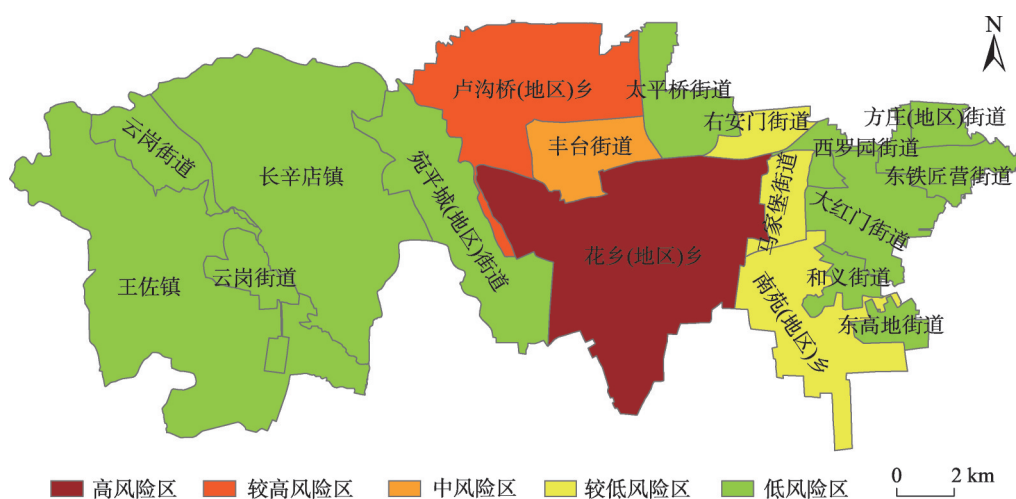


图7 丰台区乡镇(街道)风险评价结果

Fig. 7 Risk assessment results of towns (streets) in Fengtai District

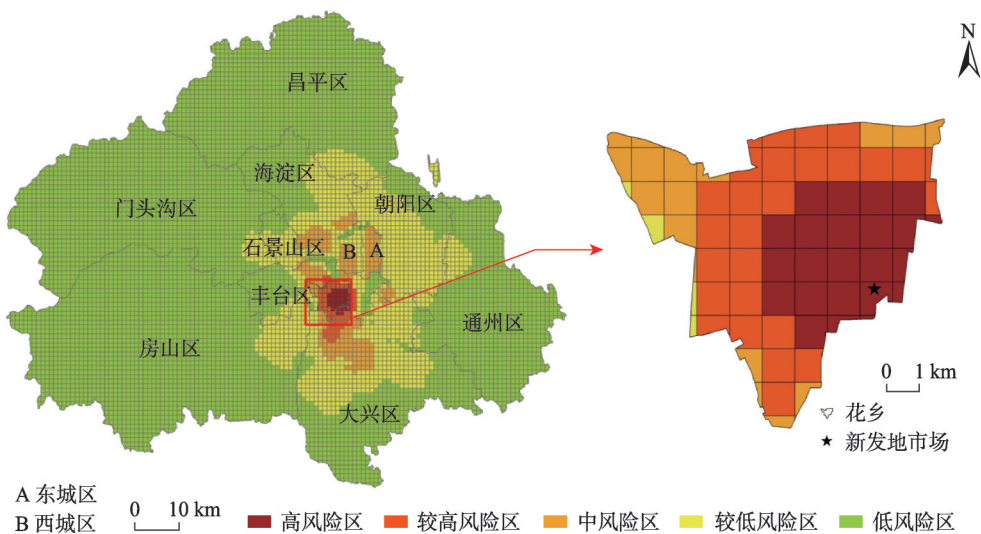


图8 1 km 格网尺度风险评价结果

Fig. 8 Results of 1 km grid scale risk assessment

区主要分布在丰台区以及大兴区,中风险区在丰台区、大兴区、海淀区、东城区、西城区、朝阳区、石景山区则有不同程度的分布,较低风险区和低风险区分布的区域最广,在所研究的北京 11 个含确诊病例区中都有所包含。这与实际的县(区)和乡镇(街道)风险评估结果具有较大的一致性。此外,基于自适应策略和乡镇(街道)风险评估结果,以花乡为例作进一步的分析。以县(区)、乡镇(街道)为评估单元,花乡所处的区域被评定为高风险区;以 1 km 格网尺度为评估单元,花乡的高风险区主要集中在新发地市场及周边 3~4 km 范围内的区域,离新发地市场较远的区域则疫情风险相对较低。相比以县(区)、乡镇(街道)为评估单元,1 km 格网尺度的风险评价结果更加详尽。

5 结论与讨论

本文提出了基于 LSTM 算法和云模型耦合的疫情传播风险预测模型,该模型融合开源时空数据,基于 GIS 和 LSTM 算法模拟精细空间尺度下疫情的空间演变过程,并在模拟结果的基础上,进一步应用云模型和自适应策略实现多空间分辨率下的疫情风险评估。以北京 2020 年 6 月份突发新冠肺炎疫情为实例,验证了模型的可行性与合理性。主要结论如下:

(1)本文提出的耦合模型从 3 个层面提升了传染病传播风险评估结果的客观性和准确性:在分析处理层,对确诊病例点要素通过核密度分析的方法较准确地表达了传染源因素在传染病传播中产生的影响,同时,精细化研究尺度(将 1 km×1 km 作为空间尺度、天为时间尺度),提高了模拟的精细程度;在模型模拟层,应用 LSTM 模型模拟疫情的时空传播过程,解决了常规机器学习模型未考虑疫情数据间时序关系的不足;在风险评估层,构建了包含传染源因素、天气因素、疫情扩散因素及疫情防御因素的疫情风险评价指标,应用云模型和自适应策略,充分考虑各区域的疫情风险状况,实现了快速定量评估不同空间分辨率的疫情风险状况。这为识别潜在疫区、针对性地制定传染病防控政策提供了强有力的技术支撑。

(2)本文依据 2020 年 6 月 11 日至 2020 年 6 月 25 日的北京 COVID-19 疫情数据,模拟了 2020 年 6 月 26 日至 2020 年 7 月 1 日疫情空间演变过程,相较

GA-BP 神经网络、决策回归树、随机森林、支持向量机模型, LSTM 模型的模拟精度更高(MAE 为 0.002 61),拟合度更好(R -square 为 0.945 5),这表明考虑疫情数据间时序关系的 LSTM 模型更适用于疫情空间演变建模。在此基础上得到的北京县(区)、乡镇(街道)和 1 km 格网 3 个空间尺度上的疫情风险评估结果与实际情况基本吻合,证实了本文所提耦合模型的有效性与可靠性。

(3)本文提出的耦合 LSTM 算法和云模型的疫情传播风险预测模型,适用于实时和短期的传染病空间传播模拟和风险评估,模型参数较少,能够较为高效、准确地评估出疫情风险,实用性和适用性较强,为模拟传染病时空传播、科学有效地评估疫情风险提供了方法借鉴。

本文提出的疫情传播风险预测模型是耦合深度学习算法和云模型,应用于传染病空间传播模拟及风险评估的初步尝试,实现了短期的疫情传播风险预测,但对于长期疫情传播风险预测的可行性还需后续研究;其次,在考虑影响新冠肺炎传播的因素上,本文未考虑输入性疫情、无症状感染者等因素的影响,随着疫情防控进入不同阶段,模型输入特征及疫情传播风险评估指标还应根据当地疫情发展特点进行调整,从而更好的为防疫工作者制定防疫措施提供技术支持;最后,本文以 1 km×1 km 为空间尺度、天为时间尺度展开疫情空间演变研究,数据粒度精细,影响因素对数据的波动性较大,如何处理异常数据、提高模型的鲁棒性将是未来改善模型性能的重要课题。

参考文献(References):

- [1] Dogan O, Tiwari S, Jabbar M A, et al. A systematic review on AI/ML approaches against COVID-19 outbreak[J]. *Complex & Intelligent Systems*, 2021,7(5):2655-2678.
- [2] 裴韬,王席,宋辞,等. COVID-19 疫情时空分析与建模研究进展[J]. *地球信息科学学报*, 2021,23(2):188-210. [Pei T, Wang X, Song C, et al. Review on spatiotemporal analysis and modeling of COVID-19 pandemic[J]. *Journal of Geo-information Science*, 2021,23(2):188-210.]
- [3] Hilker F M, Langlais M, Petrovskii S V, et al. A diffusive SI model with Allee effect and application to FIV[J]. *Mathematical Biosciences*, 2007,206(1):61-80.
- [4] Van Mieghem P. The N-intertwined SIS epidemic network model[J]. *Computing*, 2011,93(2/3/4):147-169.
- [5] Bjørnstad O N, Finkenstädt B F, Grenfell B T. Dynamics

- of measles epidemics: Estimating scaling of transmission rates using a time series sir model[J]. *Ecological Monographs*, 2002,72(2):169-184.
- [6] Das A, Dhar A, Goyal S, et al. COVID-19: Analytic results for a modified SEIR model and comparison of different intervention strategies[J]. *Chaos Solitons & Fractals*, 2021,144(2):110595.
- [7] López L, Rodó X. A modified SEIR model to predict the COVID-19 outbreak in Spain and Italy: Simulating control scenarios and multi-scale epidemics[J]. *Results in Physics*, 2021,21:103746.
- [8] 毕佳,王贤敏,胡跃译,等.一种基于改进SEIR模型的突发公共卫生事件风险动态评估与预测方法——以欧洲十国COVID-19为例[J].*地球信息科学学报*,2021,23(2):259-273. [Bi J, Wang X M, Hu Y Y, et al. A method for dynamic risk assessment and prediction of public health emergencies based on an improved SEIR model: COVID-19 in ten European countries[J]. *Journal of Geo-information Science*, 2021,23(2):259-273.]
- [9] 夏吉喆,周颖,李珍,等.城市时空大数据驱动的新型冠状病毒传播风险评估——以粤港澳大湾区为例[J].*测绘学报*,2020,49(6):671-680. [Xia J Z, Zhou Y, Li Z, et al. COVID-19 risk assessment driven by urban spatiotemporal big data: A case study of Guangdong-Hong Kong-Macao Greater Bay Area[J]. *Acta Geodaetica et Cartographica Sinica*, 2020,49(6):671-680.]
- [10] 冯明翔,方志祥,路雄博,等.交通分析区尺度上的新型冠状病毒肺炎时空扩散推估方法:以武汉市为例[J].*武汉大学学报·信息科学版*,2020,45(5):651-657,681. [Feng M X, Fang Z X, Lu X B, et al. Traffic analysis zone-based epidemic estimation approach of COVID-19 based on mobile phone data: an example of Wuhan[J/OL]. *Geomatics and Information Science of Wuhan University*, 2020,45(5):651-657,681.]
- [11] Koo J R, Cook A R, Park M, et al. Interventions to mitigate early spread of SARS-CoV-2 in Singapore: A modelling study[J]. *The Lancet Infectious Diseases*, 2020,20(6): 678-688.
- [12] Brooks L C, Farrow D C, Hyun S, et al. Flexible modelling of epidemics with an empirical Bayes framework[J]. *PLoS Computational Biology*, 2015,11(8):e1004382.
- [13] Benvenuto D, Giovanetti M, Vassallo L, et al. Application of the ARIMA model on the COVID-2019 epidemic dataset[J]. *Data in Brief*, 2020,29:105340.
- [14] Kufel T. ARIMA-based forecasting of the dynamics of confirmed Covid-19 cases for selected European countries [J]. *Equilibrium*, 2020,15(2):181-204.
- [15] Malavika B, Marimuthu S, Joy M, et al. Forecasting COVID-19 epidemic in India and high incidence states using SIR and logistic growth models[J]. *Clinical Epidemiology and Global Health*, 2021,9:26-33.
- [16] Jiang D, Hao M M, Ding F Y, et al. Mapping the transmission risk of Zika virus using machine learning models[J]. *Acta Tropica*, 2018,185:391-399.
- [17] Adamker G, Holzer T, Karakis I, et al. Prediction of Shigellosis outcomes in Israel using machine learning classifiers [J]. *Epidemiology and Infection*, 2018,146(11):1445-1451.
- [18] Chimmula V K R, Zhang L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks [J]. *Chaos, Solitons & Fractals*, 2020,135:109864.
- [19] Gao Z W, Lang M. Design of H7N9 avian influenza management and forecasting system based on GIS[C]//2015 IEEE 5th International Conference on Electronics Information and Emergency Communication. IEEE, 2015:376-379.
- [20] 任红艳,吴伟,李乔玄,等.基于反向传播神经网络模型的广东省登革热疫情预测研究[J].*中国媒介生物学及控制杂志*,2018,29(3):221-225. [Ren H Y, Wu W, Li Q X, et al. Prediction of dengue fever based on back propagation neural network model in Guangdong, China[J]. *Chinese Journal of Vector Biology and Control*, 2018,29(3):221-225.]
- [21] 李卫红,陈业滨,闻磊.基于GA-BP神经网络模型的登革热时空扩散模拟[J].*中国图象图形学报*,2015,20(7):981-991. [Li W H, Chen Y B, Wen L. Simulation of spatiotemporal diffusion of dengue fever based on the GA-BP neural network model[J]. *Journal of Image and Graphics*, 2015,20(7):981-991.]
- [22] 陈业滨,李卫红.支持向量机模型的登革热时空扩散预测[J].*测绘科学*,2017,42(2):65-70. [Chen Y B, Li W H. Simulation of spatio-temporal diffusion trend of dengue fever based on the SVM model[J]. *Science of Surveying and Mapping*, 2017,42(2):65-70.]
- [23] Pourghasemi H R, Pouyan S, Farajzadeh Z, et al. Assessment of the outbreak risk, mapping and infection behavior of COVID-19: Application of the autoregressive integrated-moving average (ARIMA) and polynomial models [J]. *PLoS One*, 2020,15(7):e0236238.
- [24] Ong J, Liu X, Rajarethinam J, et al. Mapping dengue risk in Singapore using Random Forest[J]. *PLoS Neglected Tropical Diseases*, 2018,12(6):e0006587.
- [25] Liang R R, Lu Y, Qu X S, et al. Prediction for global Afri-

- can swine fever outbreaks based on a combination of random forest algorithms and meteorological data[J]. *Transboundary and Emerging Diseases*, 2020,67(2):935-946.
- [26] 宋关福,陈勇,罗强,等.GIS 基础软件技术体系发展及展望[J].*地球信息科学学报*,2021,23(1):2-15. [Song G F, Chen Y, Luo Q A, et al. Development and prospect of GIS platform software technology system[J]. *Journal of Geo-information Science*, 2021,23(1):2-15.]
- [27] Xie Z X, Yan J. Kernel Density Estimation of traffic accidents in a network space[J]. *Computers, Environment and Urban Systems*, 2008,32(5):396-406.
- [28] 王鑫,吴际,刘超,等.基于 LSTM 循环神经网络的故障时间序列预测[J].*北京航空航天大学学报*,2018,44(4):772-784. [Wang X, Wu J, Liu C, et al. Exploring LSTM based recurrent neural network for failure time series prediction[J]. *Journal of Beijing University of Aeronautics and Astronautics*, 2018,44(4):772-784.]
- [29] Hochreiter S, Schmidhuber J. Long short-term memory [J]. *Neural Computation*, 1997,9(8):1735-1780.
- [30] Peng T, Deng H W. Comprehensive evaluation for sustainable development based on relative resource carrying capacity-a case study of Guiyang, Southwest China[J]. *Environmental Science and Pollution Research International*, 2020,27(16):20090-20103.
- [31] Byass P. Eco-epidemiological assessment of the COVID-19 epidemic in China, January-February 2020[J]. *Global Health Action*, 2020,13(1):1760490.
- [32] Paez A, Lopez F A, Menezes T, et al. A spatio-temporal analysis of the environmental correlates of COVID-19 incidence in Spain[J]. *Geographical Analysis*, 2021,53(3):397-421.
- [33] Jia J S, Lu X, Yuan Y, et al. Population flow drives spatio-temporal distribution of COVID-19 in China[J]. *Nature*, 2020,582(7812):389-394.
- [34] Zhu Y J, Xie J G, Huang F M, et al. The mediating effect of air quality on the association between human mobility and COVID-19 infection in China[J]. *Environmental Research*, 2020,189:109911.
- [35] Copernicus Climate Change Service (C3S) [DB/OL]. <https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=form>.
- [36] 潘碧麟,王江浩,葛咏,等.基于微博签到数据的成渝城市群空间结构及其城际人口流动研究[J].*地球信息科学学报*,2019,21(1):68-76. [Pan B L, Wang J H, Ge Y, et al. Spatial structure and population flow analysis in Chengdu-Chongqing urban agglomeration based on weibo check-in big data[J]. *Journal of Geo-information Science*, 2019,21(1):68-76.]