基于自增强泊松过程的 COVID-19 疫情预测

刘元浩,曹婧,沈华伟,黄俊杰,程学旗

（中国科学院计算技术研究所 数据智能系统研究中心，北京 100190）

摘要：在 COVID-19 疫情的防控工作中，对疫情传播过程中确诊人数的预测工作具有重要意义。在现有疫情传播预测工作中，以 SEIR（Susceptible-Exposed-Infected-Recovered）模型为代表的传染病模型能反映疫情相关人群人数变化，但由于其人群均匀接触的前提假设，模型的应用具有局限性。基于时间序列分析的模型可以通过简单建模历史确诊人数的时间序列对当前确诊人数进行预测，但缺乏对传染病传播的传染性、爆发性、衰减性等固有性质的认识，对疫情发展趋势变化的预测能力受到制约。为解决上述问题，该文采用基于自增强泊松过程（Reinforced Poisson Process, RPP）的模型对疫情确诊人数进行预测，考虑病毒传染性、级联传染的自增强效应和病毒传播的时效性等三个关键因子，对疫情传播的动态过程进行建模，从而对确诊人数做出预测。实验证明，相较 SEIR 模型，使用 RPP 模型进行疫情预测不依赖人群均匀混合假设，在各尺度的地理区域都有稳定且准确的预测结果，也解决了 SEIR 模型在后期预测值过高的问题；对比时间序列分析模型，RPP 模型能够掌握疫情发展的内在规律，对疫情发展前、中、后期的发展趋势预测误差分别减小 5.29%、5.04%，0.47%，并且能准确把握疫情发展的重要阶段性变化。该文方法已应用于线上平台实时疫情预测，平均误差率小于 0.5%。

关键词：COVID-19；自增强泊松过程模型；传播关键因子建模；疫情预测

中图分类号：TP3-0　　**文献标识码：A**

COVID-19 epidemic prediction based on reinforced poisson processes

LIU Yuanhao, CAO Qi, SHEN Huawei, HUANG Junjie, CHENG Xueqi

(Data Intelligence System Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: In the prevention and control of COVID-19, it is of great significance to predict the number of confirmed cases in the transmission of the epidemic. In the current works of epidemic development prediction, epidemic dynamics models, such as SEIR (Susceptible-Exposed-Infected-Recovered) model, are used to estimate the epidemic situation, but the models are limited in utilization because of the homogenous mixing hypothesis. The models based on time series analysis can predict the number of confirmed cases by simply modeling the time series of historical numbers. Unfortunately, these models also have some limitations for lacking the understanding of the inherent nature of the epidemic development, such as infectivity, explosion and attenuation. In this paper, the Reinforced Poisson Process (RPP) model were used for forecasting the number of confirmed cases by modeling three key factors of epidemic development, outbreaks of infectious virus, cascading effect of infection and aging effect of infectiousness. Experiments demonstrate that, compared with traditional SEIR model, the RPP model for epidemic prediction achieves stable and accurate prediction results in geographic areas at multiple scales, without the overestimation problem suffered by SEIR model in the later period of the epidemic development. Compared with the time series analysis method, the RPP model reduces the prediction errors of epidemic trends by 5.29%, 5.04%, and 0.47% respectively in early, middle and later period of the epidemic development, and accurately forecasts the stage changes of epidemic development. The method in this paper has

been applied to the real-time epidemic prediction on the online platform, with the average error rate less than 0.5%.

Key words： COVID-19; Reinforced Poisson Process model; modeling key factors of propagation; epidemic spread prediction

1 引言

自2019年底以来的几个月内，新型冠状病毒肺炎COVID-19在全世界范围内广泛流行，截至2020年4月7日，全球COVID-19累计确诊人数已达1,279,722例，并仍在持续快速增长。疫情的持续蔓延对人们的生命安全造成巨大威胁，也对国家医疗建设、物资调配、隔离管控等方面带来挑战。在此背景下，采用数学方法对疫情传播进行建模并对确诊病例数的增长进行及时准确地预测对于疫情防控具有重要意义。一方面，对疫情传播进行准确预测，对于医疗卫生资源的分配、防控重点的调整等具有重要的参考价值。另一方面，在防疫工作进行过程中的重要时间节点前后，对确诊病例数增长趋势的变化进行对比，能够有效地对防控措施有效性进行合理评估。

对于疫情传播预测，最常用的研究框架是传染病模型。传统的常微分方程传染病模型假设人群总数恒定且人群均匀混合[1]，通过对人群中处于各个状态的人数及各状态间的相互转换速率进行建模，推算疫情发展走势。常见的传染病模型根据人群划分的不同及人群转换的不同，包括SI[2]、SIS[3]、SIR[4]、SEIR[5]等。传染病模型从传染病传播动力学的角度进行考虑，能对疫情短期内发展趋势进行较好的模拟，但其总人数恒定且人群均匀混合的理想化假设使其应用场景受到局限。

疫情感染人数预测的另一方法是时间序列预测。疫情传播情况会随着时间推移而不断演变，疫情感染人群数可以形式化为一种时间序列，并采用时间序列分析与建模的方式进行预测。线性时间序列分析模型包括自回归（Auto-regression）模型和移动平均（Moving Average）模型[6]，以及基于两者组合而成的自回归移动平均（Auto-regression Moving Average）模型[7]。向量自回归（Vector Autoregressive）模型等非线性时间序列模型以及基于深度神经网络的RNN[8]，LSTM[9]，TCN[10]等模型也在时间序列分析问题上有优秀的表现。采用时间序列分析模型进行疫情预测，能够通过简单的模型建模疫情发展的时间序列当前值与序列历史信息间的关系，对疫情走势做出预测。但由于缺乏对疫情的传染性、爆发性、衰减性等特性的认识与建模，对疫情确诊人数的预测仍有一定的局限。此外疫情前期可用数据有限，也给时间序列模型的学习造成了很大困难。

本文采用自增强泊松过程(RPP)模型[11]对疫情确诊人数变化趋势进行预测，该模型将病毒感染人群的动态过程建模为不均匀泊松过程，通过对病毒传染性、级联传染的自增强效应和病毒传播的时效性等三个因子进行建模，对疫情传播过程中的关键因子进行刻画，以解决上述模型中出现的问题，并使用本次COVID-19疫情传播数据进行实验，证明模型的有效性。

2 相关工作

自 COVID-19 疫情发生以来，世界各地学者纷纷尝试对疫情的发展趋势展开研究和分析。其中以 SEIR 模型为代表的微分方程传染病模型占据了疫情趋势预测工作的主要部分。SEIR 模型将人群划分为易感者（Susceptible）、潜伏期感染者（Exposed）、感染者（Infected）、治愈者（Recovered）四个群体，以微分方程描述四个状态间的转换关系。2020 年 1 月 31 日，香港大学学者 Joseph T Wu 应用 SEIR 模型，利用武汉早期病例数，推测疫情会在 4 月达到高峰[12]。肖燕妮教授团队同样基于 SEIR 模型，考虑跟踪隔离等管控措施，对疫情的走势和管控举措的有效性进行了分析[13]。2 月 28 日，钟南山院士团队考虑地区间人口流动对 SEIR 模型进行改进，通过对实施管控措施时间的调整，论证了控制措施对于

基金项目：国家自然科学基金（62041207, 91746301）
作者简介：刘元浩（1998-），男（汉族），山东淄博人，中国科学院计算技术研究所博士研究生，liuyuanhao20z@ict.ac.cn；曹婧（1992-），女（汉族），浙江富阳人，助理研究员，博士，caoqi@ict.ac.cn；通讯作者：沈华伟（1982—），男（汉族），河南太康人，研究员，shenhuawei@ict.ac.cn.

减少最终 COVID-19 流行病的规模是必不可少的[14]。西安交通大学[15]、北京邮电大学[16]等国内研究机构也通过传染病动力学建模对 COVID-19 疫情走势做出了预测。上述基于传染病模型对疫情预测分析的工作被证明可以较准确地反映小范围空间在短期内的疫情走势，但由于这类模型对初始参数敏感，且基于人群均匀接触的理想假设，难以应对不同地区不同时间带来的复杂疫情发展趋势变化。

基于时间序列分析的疫情预测分析在流感等传染病预测领域多有应用[17] [18] [19]。线性时间序列模型如 Pinto[20]模型假设未来的序列值为历史序列值的线性组合，从而通过历史确诊人数对未来的确诊人数进行预测。然而线性时间序列模型在应用中面临诸多局限，基于深度学习技术的循环神经网络（Recurrent Neural Network，RNN）[8]模型解决了这一问题，但也因对长序列进行学习时会出现梯度爆炸或梯度消失现象，从而无法对长序列建模。时序卷积网络（Temporal Convolutional Networks，TCN）[10]模型通过因果空洞卷积的设计，提取序列局部特征的同时增大感受野，实现了对长时间序列的有效处理。上述提到的时间序列模型，能对时间序列进行建模并预测。然而这些时间序列模型对疫情传播的传染性、爆发性、衰减性的关键性质缺乏认识，且要求有足够的训练数据用来学习模型参数，这使得其在疫情预测应用领域存在一定局限。

3 模型方法

3.1 问题形式化

我们使用疫情传播的动态过程刻画人群中个体被病毒感染并发病这一事件的发生过程。对于某传染病 d，我们将其在时间段$[0,T]$内的疾病感染人群动态变化过程表示为个体染病事件发生的时间序列：

$$\{t_i^d\}\ (1 \le i \le n_d)$$

其中n_d表示T时刻内被疾病感染的人群总人数，t_i^d表示第i个染病事件发生的时间。不失一般性，令$0 \le t_1^d \le \cdots \le t_i^d \le \cdots \le t_{n_d}^d \le T$。

3.2 事件发生速率建模

为了建模疫情传播的动态过程中个体染病事件发生的速率，我们考察疾病传播过程中的三大现象：（1）病毒传染性，即病毒自身的传染性对最终的感染人数起决定作用；（2）级联传播所带来的自增强效应，即病毒当前的感染人数越多越容易进行新的传播感染；（3）病毒传播的时效性，即随着时间推移，病毒感染人群继续感染他人的可能性会下降。综合考虑这三个现象，我们采用自增强泊松过程（Reinforced Poisson Process，RPP）[11]来建模疾病感染人群的动态过程。具体而言，对于某个传染病 d，其感染人群的动态过程建模为一个速率为

$$x_d(t) = \lambda_d f_d(t;\theta_d) i_d(t)$$

的泊松过程。其中，λ_d是病毒自身的传染性，松弛函数$f_d(t;\theta_d)$刻画病毒传播的速率随时间演变过程。θ_d是松弛函数的参数，$i_d(t)$ 表示病毒 d 在时刻 t 已经感染的人群数量。我们假定所有的病毒在开始感染前，都有一定初始感染人数 m。因此，在第 $i-1$ 次真实感染事件发生后到第 i 次真实感染事件发生前的时间段内，我们有$i_d(t) = m + i - 1\ (1 \le i \le n_d)$。相应地，在第 n_d 次真实感染事件发生后到时刻 T 之前，我们有$i_d(t) = m + n_d$。

对于疫情预测，我们采用对数正态松弛函数

$$f_d(t;\mu_d,\sigma_d) = \frac{1}{\sqrt{2\pi}\sigma_d t}\ exp\left(-\frac{(\ln t - \mu_d)^2}{2\sigma_d^2}\right)$$

作为刻画病毒传播时效性的松弛函数。此时松弛函数的参数θ_d被替换为对数正态函数的均值μ_d和方差σ_d。

整个疾病感染人群的动态过程可以表示为如图 1 所示的产生式概率图模型。

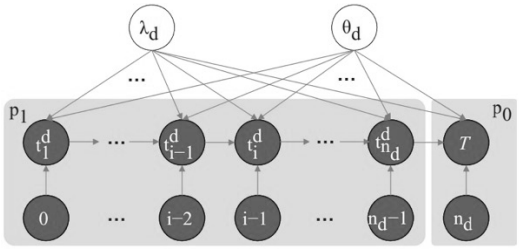

图 1 疾病感染人群动态过程的产生式概率图模型[11]

3.3 参数学习

两次连续感染事件之间的时间间隔长度服从不均匀泊松过程。因此，设第 $i-1$ 次真实感染事件的发生时刻为 t_{i-1}^d ，那么第 i 次真实感染事件在时刻 t_i^d 发生的概率满足：

$$p_1(t_i^d | t_{i-1}^d) = \lambda_d f_d(t_i^d;\mu_d,\sigma_d)(m+i-1) * e^{-\int_{t_{i-1}^d}^{t_i^d} \lambda_d f_d(t;\mu_d,\sigma_d)(m+i-1)dt}$$

在第n_d次真实感染事件发生时刻 $t_{n_d}^d$ 和观测时刻 T 之间没有感染事件发生的概率为：

$$p_0\left(T\middle|t_{n_d}^d\right)=e^{-\int_{t_{n_d}^d}^{T}\lambda_d f_d(t;\mu_d,\sigma_d)(m+n_d)dt}$$

那么，在时间间隔$[0,T]$内观测到病毒 d 的染病人群动态过程$\{t_i^d\}$的似然为

$$L(\lambda_d,\mu_d,\sigma_d)=p_0\left(T\middle|t_{n_d}^d\right)\prod_{i=1}^{n_d}p_1\left(t_i^d\middle|t_{t-1}^d\right)$$

$$=\lambda_d^{n_d}\prod_{i=1}^{n_d}(m+i-1)f_d\left(t_i^d;\mu_d,\sigma_d\right)$$

$$*\,e^{-\lambda_d\left((m+n_d)F_d(T;\mu_d,\sigma_d)-\sum_{i=1}^{n_d}F_d\left(t_i^d;\mu_d,\sigma_d\right)\right)}$$

其中，$F_d(t;\mu_d,\sigma_d)$ 是松弛函数 $f_d(t;\mu_d,\sigma_d)$的累积分布函数。

我们通过最大似然估计，学习病毒 d 的参数 λ_d，μ_d和 σ_d。令似然函数导数为零，可直接求得参数λ_d的最大似然估计值

$$\lambda_d^*=\frac{n_d}{(m+n_d)F_d(T;\mu_d^*,\sigma_d^*)-\sum_{i=1}^{n_d}F_d\left(t_i^d;\mu_d^*,\sigma_d^*\right)}$$

对于μ_d和σ_d，我们使用梯度下降法最大化似然函数，梯度

$$\frac{\partial L}{\partial\mu_d}=\frac{1}{\sigma_d}\left\{\sum_{i=1}^{n_d}\left[\tau_i^d-\lambda_d\phi\left(\tau_i^d\right)\right]+\lambda_d(n_d+m)\phi(\tau^d)\right\},$$

$$\frac{\partial L}{\partial\sigma_d}=\frac{1}{\sigma_d}\left\{\sum_{i=1}^{n_d}\left[\tau_i^d*\tau_i^d-\lambda_d\tau_i^d\phi\left(\tau_i^d\right)\right]+\lambda_d(n_d+m)\tau^d\phi(\tau^d)-n_d\right\}$$

其中，ϕ是标准正态分布的概率密度函数，$\tau_i^d\equiv\left(\ln t_i^d-\mu_d\right)/\sigma_d$，$\tau^d\equiv(\ln T-\mu_d)/\sigma_d$

3.4 疫情预测

根据泊松过程的速率函数和对应的微分方程求解，我们得到病毒感染人群的预测函数：

$$c^d(t)=(m+n_d)\,e^{\lambda_d^*\left(F_d(t;\mu_d^*,\sigma_d^*)-F_d(T;\mu_d^*,\sigma_d^*)\right)}-m$$

4 实验设置

4.1 对照模型

4.1.1 SEIR 型流行病模型

传染病学模型采用肖燕妮教授团队的工作[13]，该模型在传统 SEIR 模型对人群的“易感者-暴露者-感染者-治愈者”划分的基础上，结合 COVID-19 的实际情况与诸如检疫，隔离和治疗等干预措施，将人群分为易感者（S），暴露者（E），潜伏传染者（未表现出症状但有传染性）（A），具有症状的传染者（I），住院患者（H）和康复者（R），并进一步划分出被隔离的易感者（S_q）和被隔离的暴露者（E_q）。不同人群间的状态转换方程如下：

$$\begin{aligned}
S' &= -\left(\beta c+cq(1-\beta)\right)S(I+\theta A)+\lambda S_q,\\
E' &= \beta c(1-q)S(I+\theta A)-\sigma E,\\
I' &= \sigma\varrho E-(\delta_I+\alpha+\gamma_I)I,\\
A' &= \sigma(1-\varrho)E-\gamma_A A,\\
S_q' &= (1-\beta)cqS(I+\theta A)-\lambda S_q,\\
E_q' &= \beta cqS(I+\theta A)-\delta_q E_q,\\
H' &= \delta_I I+\delta_q E_q-(\alpha+\gamma_H)H,\\
R' &= \gamma_I I+\gamma_A A+\gamma_H H,
\end{aligned}$$

通过对模型设定合适的参数和初始值来推算疫情累计确诊人数$C=I+H+R$。

4.1.2 时间序列模型

Pinto 模型：采用该模型作为线性时间序列模型的代表。该模型划定待预测时刻前的一段时间 T 作为观测窗口，将采样窗口划分为大量的采样间隔，采用每个采样间隔内的新增确诊人数作为模型的输入，通过简单的多元线性组合给出模型的预测值[20]。

TCN 模型：非线性时间序列模型采用时序卷积网络（TCN）模型。该模型通过卷积神经网络自动提取疫情发展历史序列中的重要特征，并通过因果空洞卷积提升增大了感受野，从而可以观测更久的历史序列[10]。

4.2 实验数据

本文实验采用中国 1 月 20 日至 3 月 15 日共计 56 天的 COVID-19 每日确诊人数[21][22]作为实验数据。数据范围基本涵盖了全国自疫情开始流行至爆发到基本得到控制的全过程。同时考虑到 3 月 16 日后国内新增确诊病例来源以境外输入为主，因此将其排除，最大程度上避免了境外输入病例对实验结果的影响。

考虑到疫情传播具有地区性，不同地区疫情出现时间存在先后差异，疫情发展速度也可能不相同。我国的疫情传播呈现出明显的“武汉市-湖北省-全国其他地区”地区划分：一方面体现在地区间的隔离上，自 1 月 23 日起武汉开始全面封城，而湖北省也率先实行了较为严格的出入管制措施，最大程度上减少了感染病例的流入和流出；另一方面体现在疫情传播的时间先后和传播的规模上，国内疫情最先发现于湖北省武汉市，随后蔓延至湖北省和全国其他地区，我国及时采取措施将疫情大规模传播范围尽可能地控制在了小范围内，截至 3 月 15 日，全国近 84%的确诊病例发现于湖北省，而其中又有近 74%的病例位于武汉市。

因此本文将全国确诊人数数据划分为“全国”、“全国(除湖北)”、“湖北(除武汉)”、“武汉”四个地区层次。

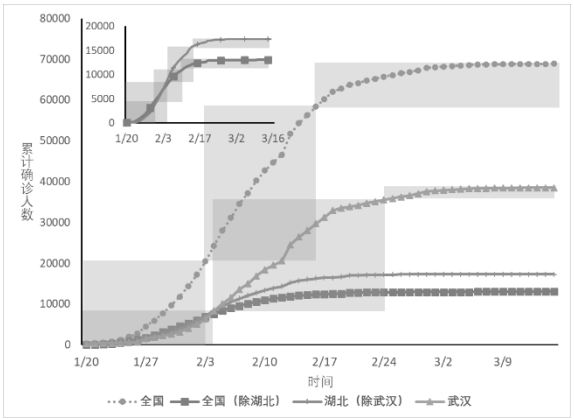

图 2 各地区确诊人数随时间变化曲线

图 2 是四个地区的累计确诊人数随时间变化的曲线（为保证模型预测结果体现疫情发展的真正趋势，我们排除掉 2 月 12 日新增确诊人数中的临床诊断病例数[22]）。不难发现各个地区划分下，疫情整体发展趋势基本一致，但存在总量和增长速度等方面的明显差异。

除了地区划分，疫情趋势在不同时间阶段的表现也有差异。如图中矩形框所标识，累计确诊人数的变化在时间上较为明显地呈现出三个阶段：（1）前期——加速增长阶段，在疫情流行初期，累计确诊人数增速持续上升，图线呈下凸经过矩形框的右下部分；（2）中期——增速稳定阶段，随着疫情发展与防控措施的实行，每日新增确诊人数基本维持不变，图线基本沿矩形对角线呈直线；（3）后期——增速放缓阶段，后期疫情得以控制，确诊人数增速迅速放缓，图线从扁平矩形框的左上部分经过。为量化表示三个阶段的特点，我们对地区a的疫情发展阶段u计算平均增长系数，

$$\overline{p_u^a} = \frac{1}{T_u}\sum_{i=1}^{T_u}\frac{c_a(i)}{c_a(i-1)}$$

其中$c_a(i)$为地区a第i天的新增确诊人数，T_u为阶段u的天数。在计算$\overline{p_u^a}$时，我们将累计确诊人数曲线进行了平滑处理以避免每日新增确诊病例数波动的影响。四个地区的各阶段划分与平均增长系数见表 1。

我们分别考察模型在各地区不同时间阶段的预测表现，作为衡量模型在不同环境下预测能力的依据。

表 1 各地区疫情发展趋势的阶段划分

地区	时间	平均增长系数
全国	$01/20-02/03$	1.34158
	$02/04-02/16$	0.97963
	$02/17-03/15$	0.83052
全国(除湖北)	$01/20-01/29$	1.34509
	$01/30-02/10$	0.94992
	$02/11-03/15$	0.86696
湖北(除武汉)	$01/20-02/04$	1.67877
	$02/05-02/13$	0.95395
	$02/14-03/04$	0.77262
武汉	$01/20-02/04$	1.57791
	$02/05-02/24$	0.98185
	$02/25-03/15$	0.80278

4.3 实验参数设置

由于确诊病例数以 1 天为单位时间统计，因此 RPP 模型的最小时间单位为 1 天，当天所有新增病例计为同时发生。由于疫情发展的情况会随时间变化，为保证模型较好地反映近期疫情的走势，我们没有使用预测时间之前的所有数据，而是在$4 \le T \le 15$范围内通过搜索确定观测窗口T大小。初始感染人数$m = 20$。

表 2 SEIR 模型参数取值

参数	取值			
	武汉	湖北(除武汉)	全国(除湖北)	全国
c	2~5	2~5	2~5	2~5
β	$2.1*10^{-8}$	$1.6*10^{-8}$	$3.2*10^{-9}$	$4.6*10^{-9}$
q	1.9×10^{-7}	1×10^{-6}	1.7×10^{-6}	1.2×10^{-6}
σ	1/7	1/7	1/7	1/7
λ	1/14	1/14	1/14	1/14
ϱ	0.86834	0.83726	0.79674	0.81285
δ_I	0.13266	0.12031	0.10282	0.11046
δ_q	0.1259	0.1302	0.1248	0.1266
γ_I	0.33029	0.01923	0.00695	0.00892
γ_A	0.13978	0.13978	0.13978	0.13978
γ_H	0.11624	0.06893	0.06972	0.07756
α	1.8×10^{-5}	2.6×10^{-4}	2.3×10^{-4}	2.5×10^{-4}

SEIR 模型参数值设定采用文献[13]中的取值，该文参数由武汉市早期疫情数据模拟获得。我们使用原文方法求取了不同地区划分下的模型参数，取值见表 2。模型的初始值获取自国家卫健委的报道数据[22],未明确报道的状态初值由预测时间前一段时间疫情相关数据通过最大似然估计得出。

时间序列模型的观测窗口大小同样通过搜索确定，从而选取合适的观测历史长度同时保证一定的训练集体量。采样间隔设置为 1 天。

4.4 评价方法

我们计算预测结果的 MAPE（Mean Absolute Percentage Error，平均绝对百分比误差）以衡量模型的预测能力。MAPE 的计算公式为

$$MAPE=\frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{\hat{C}_t-C_t}{C_t}\right|$$

其中 n 为预测时间段的天数，C_t为第t天的累计确诊人数，$\hat{C}_t$为其预测值。

5 实验结果

5.1 地区间差异对预测效果的影响

我们使用不同模型在全国、全国(除湖北)、湖北(除武汉)、武汉四个数据集上进行实验，对比不同地区间的差异对模型预测效果的影响。考虑到训练数据量和预测时段可能对各模型的预测效果产生不同的影响，因此我们分别在疫情前半段与后半段进行实验，使用 1 月 31 日之前和 2 月 10 日之前的数据训练模型，分别对随后一周（即 2 月 1 日至 2 月 7 日和 2 月 11 日至 2 月 17 日）的累计确诊人数进行预测，误差结果如表 3、表 4。

表 3 各模型 2 月 1 日至 2 月 7 日预测结果 MAPE

	RPP	Pinto	TCN	SEIR
全国	1.86%	8.44%	2.85%	108.52%
全国(除湖北)	3.62%	20.72%	4.27%	46.73%
湖北(除武汉)	3.54%	22.08%	8.25%	53.79%
武汉	6.82%	17.84%	8.22%	20.40%

表 4 各模型 2 月 11 日至 2 月 17 日预测结果 MAPE

	RPP	Pinto	TCN	SEIR
全国	2.10%	9.22%	4.16%	>1000%
全国(除湖北)	1.61%	4.09%	1.95%	474.77%
湖北(除武汉)	2.90%	5.52%	8.70%	256.27%
武汉	3.46%	3.13%	12.87%	90.33%

RPP 模型、Pinto 模型与 TCN 模型对不同地区的疫情预测效果均比较稳定。相较于 Pinto 模型与 TCN 模型仅对历史确诊人数序列进行分析，RPP 模型对疫情传播的关键因子进行了建模，其预测结果明显优于其他模型。

SEIR 模型的预测效果在不同地区差异较大。这是由于 SEIR 模型假设人群均匀混合，在全国各地采取封城措施相互隔离的情况下，绝大多数的感染人群的活动实际上被限制在了湖北省和武汉市内，这与人群均匀混合的假设高度不符，从源头上限制了 SEIR 模型的表现。

5.2 不同时间阶段对预测效果的影响

根据表 1 ，我们从时间上将疫情的发展过程划分为前期、中期和后期三个阶段。这一部分我们从时间划分的角度，考察模型“阶段内预测”和“跨阶段预测”的效果。

5.2.1　阶段内预测

疫情发展的每一个阶段都有其特定的发展趋势和规律，对这些规律的把握能力是模型完成精准预测的基本要求。这一部分实验分别使用前期，中期，后期的前半段数据作为训练数据，预测同时期后半段的累计确诊人数。以武汉地区为例，各模型预测结果如图 3、图 4、图 5。

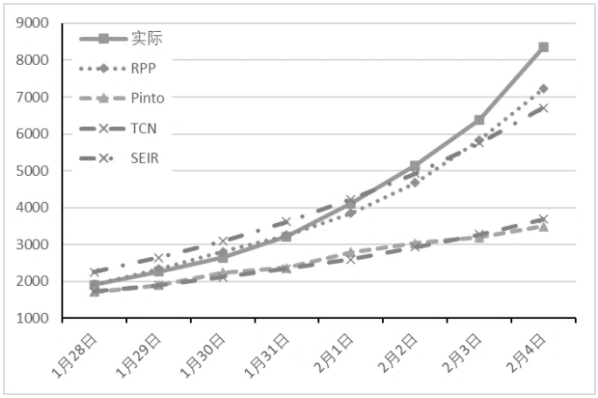

图 3 武汉市前期累计确诊人数预测结果

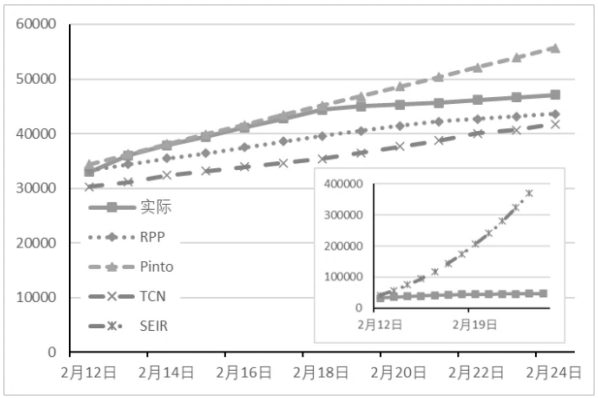

图 4 武汉市中期累计确诊人数预测结果

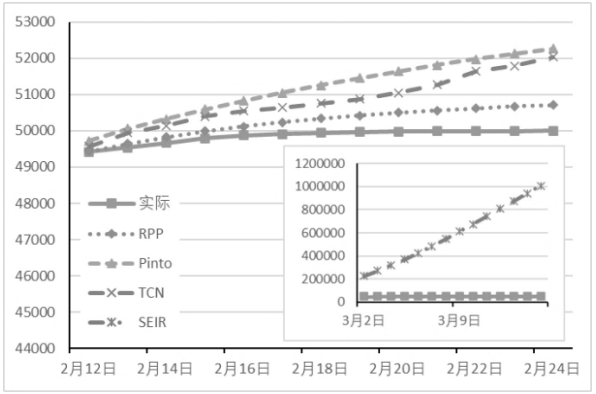

图 5 武汉市后期累计确诊人数预测结果

由图可以看出，在疫情前期数据量较少、数据规律性较弱时，Pinto 模型与 TCN 模型难以通过少量数据掌握疫情发展的整体趋势，因此预测效果较差。而 RPP 模型与 SEIR 模型通过对疫情发展固有性质的建模，可以较好地模拟确诊人数加速增长的趋势。

中后期 RPP、Pinto 与 TCN 模型对确诊人数增速保持稳定至减缓的趋势能够进行较好地模拟。SEIR 模型由于其模型假设所有人都会暴露在被传染的风险下，因此在人口总数很大时，最终累计确诊人数也会变得过高。因此我们在中后期不再对其进行对比。

我们在各地区数据集上进行了相同的阶段内预测实验，综合平均误差结果如表 5。

表 5 各模型阶段内实验累计确诊人数预测 MAPE

	RPP	Pinto	TCN	SEIR
前期	13.17%	29.23%	18.46%	12.59%
中期	4.65%	23.58%	9.69%	-
后期	0.62%	1.09%	1.77%	-

可以看出 RPP 模型在各个时间阶段内都能准确地对确诊人数进行预测，对阶段内疫情发展趋势能够进行较好地把握。

5.2.2　跨阶段预测

由于疫情发展的不同阶段趋势各不相同，从而意味着应该采取不同的防疫措施。也就是说，模型对于趋势转换的准确预测能力十分重要。因此这一部分实验分别使用前期和中期的确诊人数训练模型，预测其下一个时期的疫情趋势。

同样以武汉地区为例，各模型预测结果如图 6、图 7。

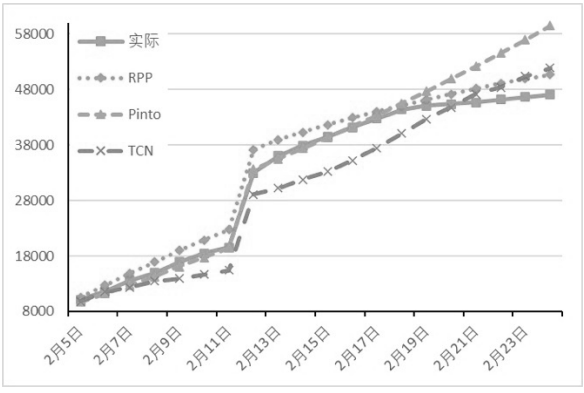

图 6 武汉市前期-中期累计确诊人数预测结果

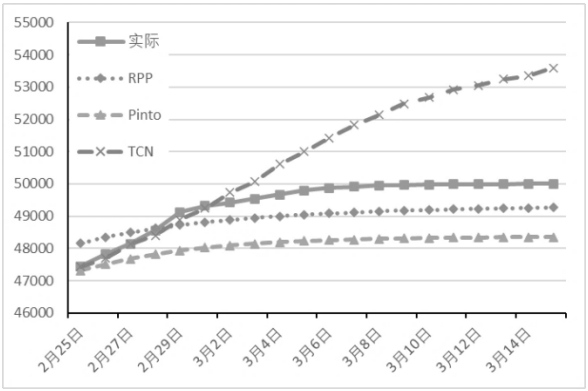

图 7 武汉市中期-后期累计确诊人数预测结果

由图 6、图 7 可知，相较于 Pinto 和 TCN 模型，RPP 模型在中期更能把握确诊人数增速保持稳定随后趋于下降的趋势。而后期 RPP 和 Pinto 模型都能较好地模拟增速迅速下降的趋势，而 TCN 模型的预测结果则倾向于保持增速持续增长。

我们在各地区数据集上进行相同的跨阶段预测实验，计算预测结果的平均增长系数(ρ̂_u^a)，再与表 1 中的实际值进行对比，计算 MAPE，结果如表 6。

表 6 各模型跨阶段实验平均增长系数的 MAPE

	RPP	Pinto	TCN
中期	6.99%	21.22%	7.78%
后期	8.83%	10.76%	28.52%

可知 RPP 模型在判断阶段变化时表现明显优于 Pinto 与 TCN 模型，在各时间阶段都能很好地预测疫情发展趋势的阶段性变化。

6 实践应用

我们将本文方法投入实际应用，自 1 月 29 日起先后对中国、日本、韩国、意大利、美国等九个国家共 12 个地区的疫情确诊人数进行预测，累计确诊人数平均误差率小于 0.5%。预测结果发布于中科天玑智疫通线上平台（https://ncov.ictbda.com/#/，效果如图 8）。

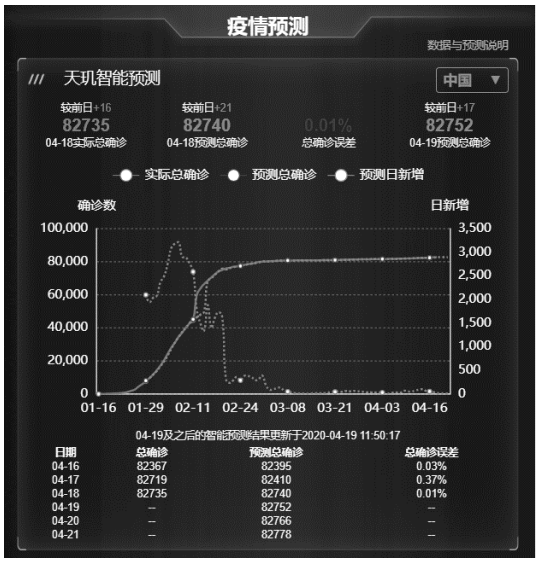

图 8 在线系统疫情预测效果

7 总结与展望

本文应用基于自增强泊松过程（RPP）的模型来预测 COVID-19 的疫情确诊病例数。我们的实验结果表明，RPP 模型在预测疫情确诊人数的任务中明显优于传统的传染病模型和时间序列分析模型。在空间上，RPP 模型克服了 SEIR 模型基于人群均匀混合的局限，在各尺度的地理区域都有稳定且准确的预测结果。在时间上，一方面，RPP 模型解决了 SEIR 模型在人口总数很大时累计确诊数持续增长的问题；另一方面，RPP 模型通过建模疫情发展过程中的关键因素，摆脱了时间序列分析模型仅对历史数据建模的局限性，从而对疫情发展各个阶段的疫情走势能够进行更精确的预测，并且能准确把握疫情发展的重要阶段性变化，其结果在实际应用更具有参考价值。

本文的方法也存在进一步优化的空间，本文假设感染速率与当前感染人数成正比，并使用松弛函数从整体上描述部分感染者被隔离或被治愈等情况造成的感染者总体影响力下降。未来将考虑使用 Hawkes 过程进行建模，细化不同状态感染者对疾病感染速率的影响。

参考文献

[1] Barabsi A L. Network Science[M]. New York: Cambridge University Press, 2016: 378-388

[2] Maureen Hurley, Glen Jacobs, Melinda Gilbert. The basic SI model[J]. Special Issue:Supplemental Instruction: New Visions for Empowering Student Learning, 2006(106):11-22.

[3] Hethcote H W, Yorke J A. Gonorrhea transmission dynamics and control[J]. Lecture Notes in Biomathematics ,1984, 56.

[4] Kermack W O, McKendrick A G. A contribution to the mathematical theory of epidemics[J]. Proceedings of Royal Society of London, 1927, 115(772):700-721.

[5] Greenhalgh D. Hopf bifurcation in epidemic models with a latent period and nonpermanent immunity[J]. Mathematical & Computer Modelling, 1997, 25(2):85-107.

[6] Hipel K W, McLeod A I. Time series modelling of water resources and environmental systems[M]. Elsevier, 1994: 91-102.

[7] Box GEP, Jenkins GM, Reinsel G C, et al. Time Series Analysis, Forecasting and Control(5th Edition)[M]. New Jersey: John Wiley and Sons Inc, 2015:712.

[8] Goodfellow I, Bengio Y, Courville A．Deep learning (Vol. 1) [M]. Cambridge: The MIT Press,2016:367-415.

[9] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural computation, 1997, 9(8):1735-1780.

[10] Bai S, Kolter J Z, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling [EB/OL]. https://arxiv.org/abs/1803.01271?spm=a2c4e.11153940.blogcont642474.26.18fb6a59FjzYuR ,2018.

[11] Shen H , Wang D , Song C, et al. Modeling and predicting popularity dynamics via reinforced poisson processes [C].Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI'14),Canada: 2014: 291–97.

[12] Wu J T, Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study[J].The Lancet, 2020, 395(10225): 689-697.

[13] Tang B, Wang X, Li Q, et al. Estimation of the transmission risk of 2019-nCov and its implication for public health interventions[J]. Journal of Clinical Medicine, 2020, 9(2): 462.

[14] Yang Z, Zeng Z, Wang K, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions[J]. Journal of Thoracic Disease, 2020, 12(3): 165–174.

[15] Cao S, Feng P, Shi P. Study on the epidemic development of corona virus disease-19 (COVID-19) in Hubei province by a modified SEIR model[J]. J Zhejiang Univ (Med Sci), 2020, 49(2): 178-184.

[16] Zhang Lin. Fitness of the generalized growth to the COVID-19 data[J]. Journal of University of Electronic Science and Technology of China,2020,49(3): 345-348

[17] Hu Y, Liao J, Feng G, et al. Application of multiple seasonal autoregressive integrated moving average model in prediction of incidence of hand foot and mouth disease in China[J]. Disease Surveillance, 2014, 29(10): 827-832

[18] 彭志行,陶红,贾成梅,等.时间序列分析在麻疹疫情预测预警中的应用研究[J].中国卫生统计,2010,27(05):459-463.

[19] 沈冰,杨晓明,卑伟慧,等.时间序列分析在上海静安区流感样病例预测预警中的应用[J].环境与职业医学,2016,33(02):156-159.

[20] Pinto H , Almeida J M , Gonçalves M A. Using early view patterns to predict the popularity of youtube videos[C]. WSDM '13: Proceedings of the sixth ACM international conference on Web search and data mining , Rome Italy, 2013.

[21] 湖北省卫生健康委员会.防控新型冠状病毒感染肺炎疫情信息发布[EB]. http://wjw.hubei.gov.cn/bmdt/ztzl/fkxxgzbdgrfyyq/xxfb/.

[22] 中华人民共和国国家卫生健康委员会. 新型冠状病毒肺炎疫情防控疫情通报[EB]. http://www.nhc.gov.cn/xcs/yqtb/list_gzbd.shtml.