新冠疫情时空数据的自动采集与分析方法

段　裕

学　　院：计算机科学与技术学院　　**专　　业**：计算机科学与技术

学　　号：180110704　　**指导教师**：冯山山

2022 年 6 月

哈尔滨工业大学深圳校区

毕业设计（论文）

题　　目　新冠疫情时空数据的自动采集与分析方法

姓　　名　段裕

学　　号　180110704

学　　院　计算机科学与技术学院

专　　业　计算机科学与技术专业

指 导 教 师　冯山山

答 辩 日 期　2022 年 06 月 09 日

摘　要

2019 年末新型冠状病毒肺炎疫情爆发，随后迅速蔓延扩散并造成大范围的感染。2020 年 3 月，世界卫生健康组织宣布新冠疫情构成世界大流行。新冠疫情对人们的生产生活和身心健康造成了巨大的冲击和危害，展开对中国新冠疫情时空数据的收集工作以及基于模型的研究十分必要。

本项目收集和整理了自 2020 年 01 月至 2022 年 03 月期间中国各省市的新冠疫情数据，包括中国各省市每日新冠肺炎确诊情况、中国各级行政区划基本信息等，此外还收集了深圳市病例的详细信息。本项目同时展开了数据集可视化工作，设计实现了中国疫情实时可视化系统，该系统从不同角度以多种图表形式对中国各省市的疫情实时数据进行了可视化，包括全国新冠肺炎确诊数量的累计趋势和新增趋势折线图、中国疫情实时地图等。

本项目还从不同的角度对数据集进行了建模分析。首先从传统的流行病动力学角度进行建模，使用到了 SEIR 系列模型；接着考虑数据的时序特点，从长短期记忆网络的角度进行了研究；最后进一步从时空点过程的角度进行了实验分析。实验结果表明，新冠疫情的发展趋势总体较为符合 SEIR 系列模型的假设，可以通过 SEIR 系列模型来模拟和预测疫情爆发初期的发展趋势；在训练样本数据充足时，长短期记忆网络模型能够达到更优的性能，但在训练样本不足时，长短期记忆网络无法有效的学习到疫情发展的趋势；通过条件强度函数能够完全地描述一个时空点过程模型，时空点过程分析法适合用于处理疫情时空数据。

关键词：新型冠状病毒；流行病动力学模型；长短期记忆网络；时空点过程；疫情数据可视化；预测模型

Abstract

At the end of 2019, the epidemic of novel coronavirus pneumonia broke out, and then spread rapidly and caused widespread infection. In March, 2020, the World Health Organization announced that the COVID-19 has become a world pandemic. The COVID-19 caused great impact and harm to people's production, life and physical and mental health, thus it is necessary to collect the spatio-temporal data of COVID-19 in China and conduct model-based research.

This project has collected and sorted out the data of COVID-19 in various provinces and cities in China from January 2020 to March 2022, including the daily diagnosis of COVID-19 in various provinces and cities in China, the basic information of administrative divisions at all levels in China, and also collected the detailed information of cases in Shenzhen. At the same time, the project has carried out data set visualization, designed and implemented a real-time visualization system for China's epidemic situation. The system visualizes the real-time epidemic situation data of various provinces and cities in China from different angles in a variety of chart forms, including the cumulative trend and new trend line chart of the number of confirmed COVID-19 in China, and the real-time epidemic situation map of China.

The project also conducted modeling and analysis of data sets from different perspectives. Firstly, the traditional epidemic dynamics model is built, and SEIR series models are used; Then, considering the time series characteristics of data, the research is carried out from the perspective of short-term and long-term memory network; Finally, the experimental analysis is carried out from the point of view of spatiotemporal point process. The experimental results show that the development trend of COVID-19 is generally in line with the hypothesis of SEIR series models, and SEIR series models can be used to simulate and predict the development trend at the initial stage of the outbreak; When the training sample data is sufficient, the long-term and short-term memory network model can achieve better performance, but when the training sample is insufficient, the long-term and short-term memory network cannot effectively learn the trend of epidemic development; The conditional intensity function can completely describe a spatiotemporal point process model, and the spatiotemporal point process analysis method is suitable for processing spatiotemporal data of epidemic.

Keywords: novel coronavirus，epidemic dynamics model，
long-short-term memory network，spatial-temporal point process
visualization of epidemic data，prediction model

目 录

摘 要 ……I
ABSTRACT ……II
第 1 章 绪 论……1
1.1 课题背景及研究的目的和意义……1
1.2 疫情趋势预测模型及其相关理论的发展概况……1
1.2.1 传染病动力学模型分类……1
1.2.2 经典仓室模型及相关研究工作的发展……2
1.2.3 LSTM 模型及相关研究工作的发展……3
1.3 时空点过程模型的发展概况及其在流行病学中的应用……3
1.3.1 时空点过程模型分析法概述……4
1.3.2 点过程模型的分类与例述……4
1.3.3 时空点过程模型在流行病时空传播分析的应用……5
1.4 本文的主要研究内容……6
1.5 本文的组织结构……6
第 2 章 新冠疫情时空数据的获取与可视化 ……7
2.1 引言……7
2.2 网络爬虫的一般概念与工作流程……7
2.2.1 爬虫技术的概念与工作流程……7
2.2.2 数据爬取预处理系统的模型建立……7
2.3 新冠疫情数据集的获取……9
2.3.1 新冠疫情数据集总体情况说明……9
2.3.2 省市级疫情实时数据的爬取……10
2.3.3 深圳市卫健委病例数据的爬取……12
2.3.4 中国各级行政区划信息的爬取……13
2.4 新冠疫情数据可视化系统……14
2.4.1 中国实时疫情信息的可视化分析……14
2.4.2 深圳市疫情信息的可视化分析……15
2.5 本章小结……17
第 3 章 新冠疫情数据的预测分析方法 ……18
3.1 引言……18

3.2 SEIR 系列模型原理详述……18
3.3 基于深圳市数据的 SEIR 系列模型分析……21
3.4 基于长短期记忆网络的时序分析……23
3.4.1 RNN 网络模型的工作原理和存在的问题……24
3.4.2 LSTM 网络的工作原理与改进点……24
3.4.3 使用 LSTM 网络分析疫情数据……25
3.5 模型间的比较分析……26
3.6 本章小结……27
第 4 章 基于时空点过程的新冠疫情数据分析 ……28
4.1 引言……28
4.2 时空点过程的事件建模法……28
4.2.1 时空点过程的一般形式……28
4.2.2 几种点过程模型的探讨……29
4.3 基于点过程模型的新冠疫情数据分析……31
4.3.1 数据的预处理……31
4.3.2 点过程模型的部署与训练……31
4.4 本章小结……32
结 论 ……33
参考文献……34
原创性声明……36
致 谢……37

第 1 章　绪　论

1.1　课题背景及研究的目的和意义

2019 年 12 月春节期间，中国湖北省武汉市报告了首例新冠肺炎病例，在春运高峰期等因素地加持下，新冠疫情失控并迅速开始流行。随后，中国总体上采取遏制的方针，较好地遏制了国内本土疫情，与此同时，其它国家和地区也出现了新冠肺炎病例并迅速蔓延。日内瓦时间 2020 年 3 月 11 日，世界卫生组织宣布本次疫情构成全球大流行，国际疫情仍处于爆发态势。

在如今经济全球化的时代，货物、人口流动频繁，人类社会经济文化空前发展，但与此同时这也加剧了各种传染病的产生、爆发和流行的可能。因此，在新冠疫情逐渐常态化的今天，展开对中国新冠疫情数据的收集工作，以及利用数学建模方法与机器学习模型对新冠疫情传播过程进行描述、分析、预报和控制具有重要的意义。同时，国家卫生健康委员会、地方政府会对每日的新冠疫情情况进行通报，部分网络平台会依据上述通报上线实时播报平台，这是本文研究的数据基础。

1.2　疫情趋势预测模型及其相关理论的发展概况

从系统科学的角度来看，传染病流行是在社会人群中发生的一个复杂扩散过程，针对这一扩散过程对传染病传播流行机理进行分析表达的理论方法即为传染病动力学模型[1]。另一方面，整理得到的疫情数据中包含时序数据，可以使用 LSTM 网络进行一定的分析。以模型为基础对传染病进行分析和预测，有助于理解传染病的流行机理，为人工干预措施的选择提供理论依据。

1.2.1　传染病动力学模型分类

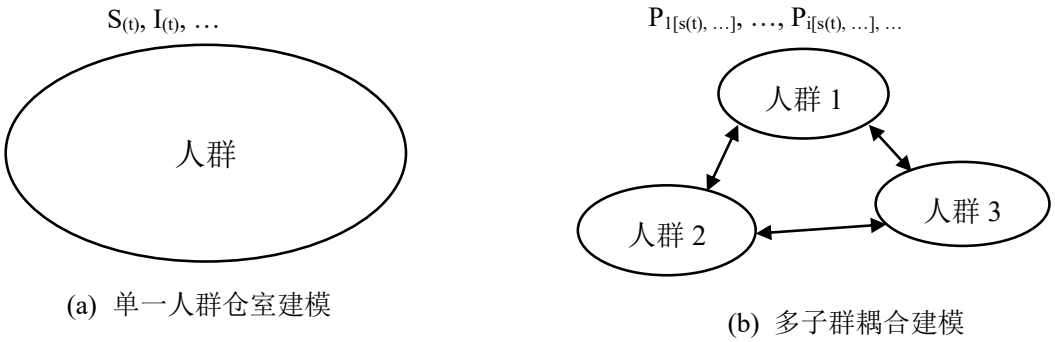
人们对于传染病动力学模型的研究已经有相当长的历史，其间出现了许多优秀的建模范例，在机器学习兴起之前，这些建模理论方法往往建立在数学分析的基础上并通过（偏）微分方程组进行表达。这些模型各有所长，适用于不同的案例分析，从研究对象的划分这一角度分类，大致可以分为三类[1]：

(a) **单一人群仓室建模**　模型将整个人群视为一个整体，并把这个整体放到一个虚拟的封闭仓室中研究，在人群内部进行类型划分，流行过程表现在易感者,感染者等各类人员集计量的变化。

(b) **多子群耦合建模** 在一个大型的人群中，人群存在空间聚集性，由此形成了不同的子群，子群内部和子群之间均有人员流动，模型以子群作为研究对象。

(c) **微观个体网络建模** 该模型把每个个体视为研究对象，个体包含各种属性和行为，不同的个体属性不尽相同，个体之间的接触关系构成了网络上的传播过程。

三种建模方法的基本思想如图 1-1 所示：



(a) 单一人群仓室建模

(b) 多子群耦合建模

(c) 微观个体网络建模

图 1-1　传染病动力学模型结构的分类

图 1-1 中展示了三种传染病传播模型的核心思想。图(a)是单一人群仓室模型，将人群看作一个整体，直接划分为S（易感者）、I（感染者）等仓室，由于仓室人数随时间变化，因此分别记为$S(t)$、$I(t)$等。图(b)是多子群耦合模型，一个实际的庞大的群体往往存在空间局部聚集性，模型据此将人群划分为多个子群$P_1[s(t),\dots],\dots,P_i[s(t),\dots]$，在子群内部进行不同仓室的划分。图(c)表示微观个体网络模型，每一个个体$_{i=1}^{n}P_i$包含不同的属性$_{i=1}^{n}s_i$，不同属性的个体使用不同颜色表示，个体间的连线表示两个个体发生了接触联系，因而微观个体网络模型研究的对象更细致，考虑的因素也更加复杂。

1.2.2　经典仓室模型及相关研究工作的发展

经典仓室模型均属于单一人群模型的范畴，模型把整个人群视为一个整体，人群因患病情况不同而划分为不同的仓室，模型通过动力学方程描述不同仓室的人员数量变化。最经典的仓室模型（compartmental model）是 SIR 模型，由 Kermack 等于 1927 年提出[6]，正如模型名称所示，该模型把人群划分为三个仓室：易感者（Susceptible，S）、感染者（Infected，I）、康复者(Recovered，R)。

在 SIR 模型的基础上，为了适应不同流行疾病的爆发趋势，仓室模型主要在以下几个方面进行了扩展[1]：一是采用不同的仓室设置；二是考虑人口动力学，特别是将人口年龄结构看作重要因素；三是考虑更多的因素，如随机性、人口、空间的异质性等。

考虑其中针对仓室设置的扩展，有如下说明[1]：

仓室对应着传染病流行过程中的可能状态，仓室的划分首先与传染病特性有关。经典 SIR 模型适合治愈后获得终生免疫的疾病。某些疾病治愈后不能获得免疫力，则可用 SIS 模型。若康复后能够获得一定的免疫力，但免疫力会逐渐消失，则可用 SIRS 模型。某些疾病有潜伏期（暴露，exposed）， 则可用 SEIR 模型。某些疾病新生儿可从母亲获得被动免疫,但-段时间后免疫力消失，则可采用 MSEIR 模型（M 表示从母亲获得被动免疫的人群）。与此类似，还有许多仓室设置方式，一般根据仓室和转移路线称为 SI、SEI、SEIRS、MSEIRS 等。

1.2.3 LSTM 模型在疫情数据上的分析

循环神经网络（Recurrent Neural Network， RNN）采用循环链式结构处理问题输入，其中的循环重复节点称为闭合隐藏中间单元。RNN 的闭合隐藏神经单元结构较为简单，在数据处理的过程中会出现累乘算式，容易引起梯度消失或梯度爆炸问题[3]。长短期记忆网络（Long Short Term Memory network, LSTM）是一种特殊的 RNN，相对于 RNN，其闭合隐藏中间单元更加复杂，可以很好地解决长时依赖问题[7][8]。具体的，LSTM 主要通过引入门控机制来控制信息的累积速度,包括有选择地加入新的信息,并有选择地遗忘之前累积的信息[4]，由此缓解长序列训练过程中的梯度消失和梯度爆炸问题。

新冠疫情数据中包含时序性的数据序列，可使用 LSTM 模型进行分析。赵永翼等人针对 2020 年 2 月 22 日—7 月 13 日的新冠疫情新增数据等进行了序列分析预测[6]，实验基于 python 的 keras 框架和 tensorflow 深度学习库进行，根据前 3 天预测第 4 天累计确诊人数。实验表明 LSTM 网络适合用于做疫情序列分析预测。

1.3 时空点过程模型的发展概况及其在流行病学中的应用

1.3.1 时空点过程模型分析法概述

点过程的概念本身就包含着时空的因素，点是指定义在有界连续的时间域或空间域（例如一个欧几里得平面或三维的欧几里得空间等）中的事件（events），而过程（process）则蕴含着时间推进的思想，因而点过程可以更也称为时空点过程。

时空点过程建模分析法能对众多真实场景中产生的数据建模。例如在 1978 年 Ogata 教授等人基于时空点过程研究了地震爆发的位置和震级数据，预测余震发生的位置和时间，论文采的预测方法是通过仿真来生成未来的余震事件，并提出了一种基于 Shedler-Lewis 细化算法（Shedler-Lewis thinning algorithm）的改进细化算法[6]。Stoyan 教授将时空点过程建模方法应用于林业，对森林科学数据展开分析，指出点过程的理论统计方法和随机模型的快速发展在某种程度上完全是受林业应用的启发，特别是现代已开发的单树模型[7]。

地震数据、森林数据以及疫情数据等都是多维异步事件数据，它们互相影响并在连续时间域上呈现出复杂的动态规律。基于时空点过程的序列分析法，一个重要的应用就是对未来的事件进行预测。图 1-2 给出了多维异步事件序列的一个示意图，三个用户的行为被表征为不同的事件序列，不同类型的事件由不同形状表征，虚线箭头代表事件间前后关联。

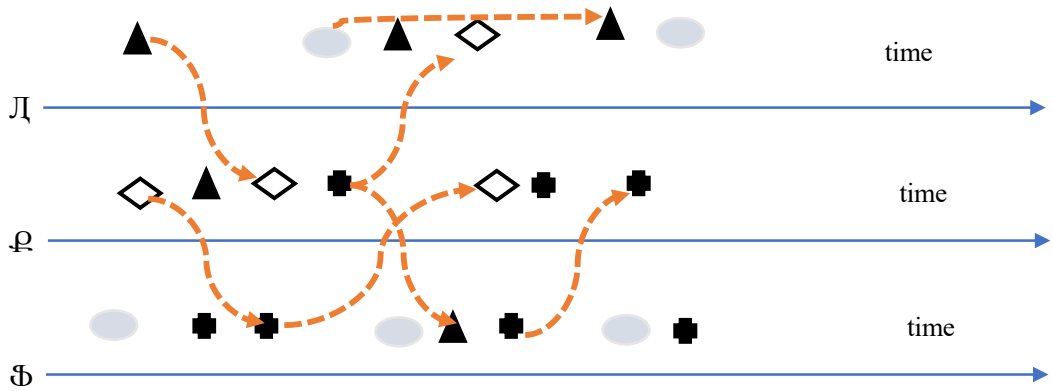


图 1-2　类别标记的连续时间域异步事件序列及其相互影响

图 1-2 中，箭头（激励关系）都是指向时间轴的正方向，表明历史事件会影响未来事件的时空过程特征。同时，一部分箭头从一个用户的事件指向另一个用户事件，另一部分则从一个用户的事件指向同一个用户的另一个事件，这体现着时空点过程研究对象（事件）的多维异步性。

1.3.2 点过程模型的分类与例述

点过程，属于随机过程的范畴，以空间泊松过程为例，假定研究对象是一片大草原，在这一片草原上随机选定一个样方（选区），要求统计某种植物或昆虫的数量，统计结果记为样方数，并称这样一次操作为抽样，假定每次抽样相互独立，那么样方数就服从空间泊松过程。传统的点过程模型使用数学函数描述点过程的强度，具备一定的数学物理意义，而随着机器学习的兴起，点过程模型逐渐转向使用深度网络学习模拟强度函数。典型的点过程模型包括[8]：

(a) **泊松过程（Poisson process）**：泊松过程把强度函数视为与历史事件无关的量，其中齐次泊松过程认为强度函数为一个恒定的常量，而非齐次泊松过程则认为 λ(t)是随时间变化的函数。Cox 教授提出了经典的重随机泊松过程[9]，将强度函数看成是一个受到各种外部因素影响的随机变量。

(b) **自校正过程（self-correcting process）**：自校正过程把把强度函数视为与历史事件无关的量，在没有新事件发生时，强度会持续上升，当一个事件发生后，强度函数会以指数的形式回落[8]。

(c) **自/互激励过程（self/mutual-exciting process）**：自激励过程，亦称为霍克斯过程（Hawkes processes）[10]，其内部机制表示发生的历史事件对于未来事件的发生有激励作用，并且历史事件的影响以累加的形式进行叠加。

1.3.3 时空点过程模型在流行病时空传播分析的应用

在 2012 年 Meyer 教授等发表的《基于烈性脑膜炎球菌病流行过程的条件强度时空模型分析》一文中[11]，Meyer 教授及其团队提出了基于条件强度函数（conditional intensity function）的时空点过程模型，它提供了一个参数框架，用于在疾病监测中典型的时空点过程中进行前瞻性变化点分析：在随机过程控制的框架内，例如，可以使用似然比检测器来监控时间需要包含流行病成分来描述观察到的数据的点。这在想法上与同质时空泊松过程设置相对应。

Schoenberg 教授使用 2014 西非埃博拉疫情的数据来评估简单的霍克斯点过程模型在预测埃博拉病毒在几内亚、塞拉利昂和利比里亚的传播预测的效果[12]。为了比较，使用 SEIR 模型在相同数据集上使用相同的指标进行评估。为了测试每个模型的预测能力，论文通过使用前 75%的数据进行估计和随后的 25%的数据进行评估来模拟在实际爆发期间进行近乎实时的预测的能力。

RTQ Chen 教授在研究神经点过程的相关工作中[13]，使用美国新泽西州自 2020 年 03 月 15 日至 2020 年 08 月 01 日共 139 日的数据，并从时空点过程分析的视角对多个模型泊松过程、自校正过程、霍克斯过程、神经霍克斯过程等等进行了研究分析。

1.4 本文的主要研究内容

本文对几类流行病动力学模型包括 SEIR 模型以及改进的 SEIR 模型进行了模拟和仿真实验，结合深圳市早期疫情爆发时的历史疫情数据集对 SEIR 模型与改进 SEIR 模型进行了参数拟合。同时，结合最新研究的长短期记忆网络在与 SEIR 系列模型相同的数据集上进行了训练和预测，结合均方根误差（RMSE）和平均绝对误差（MAE）等误差标准对 SEIR 系列模型以及 LSTM 模型的预测结果进行了评估。结果表明，在训练样本容量足够时，LSTM 模型的预测结果最好，预测值与真实值相当吻合，改进的 SEIR 模型预测结果其次，SEIR 模型预测结果误差略大。此外，本文还从点过程分析法的视角，对疫情的扩散过程进行了研究，并采用 RTQ Chen 教授在研究时空点过程中使用的方法[13]，对中国各省市历史新冠确诊数据集进行了研究和分析。此外，针对前期工作中收集整理得到的数据集，本文从不同的角度对其进行了描绘和可视化。

1.5 本文的组织结构

本文一共分为四章对研究内容进行展开叙述。

第一章为绪论：首先对本文的研究对象中国新冠疫情时空数据、研究目的和意义进行了简要概括。其次对本文将要使用到的模型及相关研究工作进行了叙述和说明，包括仓室系列模型以及 LSTM 网络和时空点过程模型等的相关理论。

第二章为新冠疫情时空数据的获取与可视化，详细说明了中国新冠疫情时空数据、深圳市患者个案数据以及中国各地级市人口、面积、经纬度等相关数据的获取方法。同时介绍了基于已有的数据集进行的多方面可视化工作，完成了以 Python 与 Flask+MySQL 为基本框架的展示系统。

第三章为新冠疫情数据的预测分析方法，详细阐释了 SEIR 系列模型和 LSTM 模型的原理，并对其中部分模型进行了仿真，同时以深圳市早期疫情数据集为基础，对 SEIR 模型、mSEIR 模型和 LSTM 模型进行了训练，将各自的预测结果进行误差分析和比较。

第四章为从时空点过程的角度分析疫情数据，针对 RTQ Chen 教授等人研究的时空点过程方法对中国历史疫情数据进行了分析。主要完成了泊松过程、自校正过程和霍克斯过程的模型训练，阐述了极大似然估计的方法原理，利用对数似然值对结果进行了探讨分析。

第 2 章　新冠疫情时空数据的获取与可视化

2.1　引言

真实可信的数据是后续展开建模分析与可视化工作的基础。目标数据往往散布在离散的网页上，而网页是一种半结构化的数据结构，可以包含文字、图片、音视频等等信息，因而其内容往往是存在大量冗余的，需要进行过滤和预处理。一般而言过滤和预处理过程是多层次的，其最终目的是得到程序可处理的数据输入。这一系列工作即定义了本文使用的数据爬取预处理系统，另一方面，为了更好的展示实时爬取的数据，本文还设计实现了疫情实时数据可视化系统。

2.2　网络爬虫的一般概念与工作流程

2.2.1　爬虫技术的概念与工作流程

作为搜索引擎的基础构件之一，网络爬虫 (Crawler)直接面向互联网，它是搜索引擎的数据来源，决定着整个系统的内容是否丰富 、信息能否得到及时更新 。它的性能表现直接影响整个搜索引擎的效果。一个高性能的 Crawler 需要从以下四个方面来考虑[14]：

首先爬虫必须满足可伸缩性，爬虫能通过增加硬件资源使性能得到线性提高。其次爬虫是分布式的，必须支持分布式的爬行以满足目前互联网的规模。然后爬虫必须有限制爬行，爬虫不能在短时间内大数据量地集中访问同一个主机下的网页，否则会影响普通用户对其的访问，进而可能被对方限制访问。最后是可定制性，可根据不同的爬行任务和特定的主题定制相应的功能模块，使功能插件化，打造个性化爬虫。

在本文中，给出一个可能的定义：网络爬虫(Web Crawler)是从指定的初始 url 队列开始，对于每一个队列中的 url，通过网络资源请求获取到网络页面，并结合网页解析技术提取有效信息同时把过滤得到的新 url 添加到初始 url 队列尾并继续爬取页面的自动化过程，这一过程在达到用户指定的结束条件时终止。

2.2.2　数据爬取预处理系统的模型建立

一个良好运作的系统一般要满足功能模块化层次化、高内聚低耦合的特点，本文设计的数据爬取预处理系统将数据提取工作细化，并分别交给爬取模块、数据预

处理模块和数据存取模块。这三个模块的说明如下：

（1） 爬取模块

爬取模块负责爬取原始网页，将网页的目标内容存入文档，是数据爬取预处理系统的关键模块。需要获取的目标数据属于文本数据，因而文档格式采用逗号分隔文件（csv 文件），优点是文档属于纯文本文档，占用空间小，同时又可以借助 python 的 pandas 库进行存取，操作类似处理表格时按行按列操作。

爬取模块地核心网络爬虫程序，程序必须按照 2.2.1 中对爬虫的定义严格执行，同时按照需求尽可能满足高性能的 Crawler 的特性：

首先对于可扩展性与分布式的要求。由于本统仅针对几个特定的平台或发布页进行信息获取，数据量并不大，初始 url 队列较小。因此，可以暂时不考虑这两个层次的要求。

其次，爬虫必须有限制地爬行，否则可能面临拒绝访问的风险。同时与通常的程序不同，爬虫程序是一种网络程序，需要考虑网络的波动性与服务器的负载能力，例如，需要编写异常处理模块以应对请求页面超时等问题，需要在连续获取页面时插入一段随机等待时间，需要为网络请求函数封装请求头部，并随机更换等等。

最后针对可定制性的要求，在使用 Python 语言编写爬虫模块的背景下，可以将其封装为一个类，例如仅针对类似发布页信息爬取的爬虫类，它可以自动的翻页和获取文章列表，同时获取文章具体内容。

（2） 数据预处理模块

数据预处理模块负责按照模型训练的数据输入要求对爬取得到的 csv 文件中的原始数据进行过滤和转换，是数据爬取预处理系统的重要模块。这一步往往是多层次的，需要一步步的过滤以及转化，其中过滤过程用到的主要方法是直接字符串匹配法、字符串正则匹配法等；转化过程一方面需要将一些非结构化的信息转为结构化的信息，例如对于提取出的地点信息，应当结合有关地图 API 得到经纬度的位置信息，另一方面则需要进行一定的计算处理，例如对于日期信息，应该确定起止日期，并将起止日期以及之间的所有日期转为距起始日期的时间间隔等。

（3） 数据存取模块

数据存取模块负责将处理好的数据存入数据库进行固化，本文采用 MySQL 数据库，借助 MySQL 脚本进行数据存取，因此这一部分的工作主要交由 MySQL 数据服务器处理。

如图 2-1 展示了数据爬取预处理系统的主要工作流程：

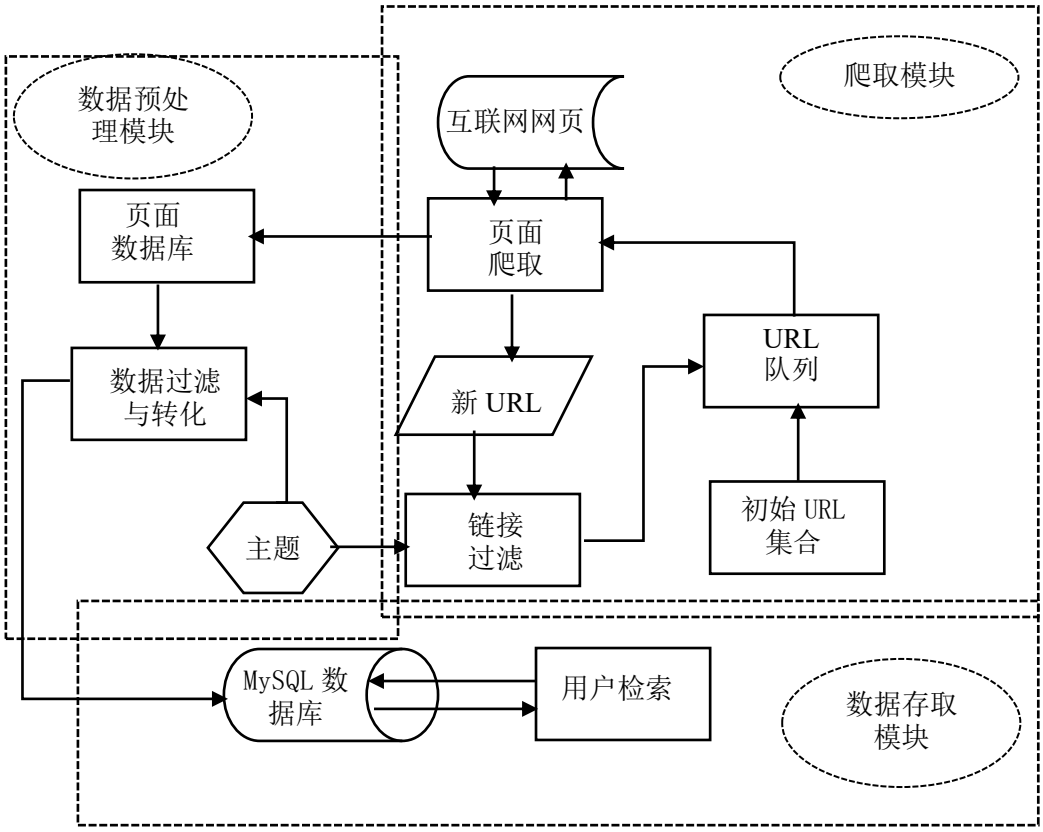


图 2-1　数据爬取预处理系统工作流程

图 2-1 中，整个系统的工作流程被划分为三个部分：

其一是爬取模块，模块维持一个 URL 队列，队列初始化为初始 URL 集合，此后爬虫程序每次从 URL 队列中取一个 URL 并访问得到页面，一方面将页面数据交给数据预处理模块，另一方面依据主题过滤得到页面中的目标 URL 并放入队列。

其二是数据预处理模块，模块将接收到的页面数据依据主题进行过滤和转化得到适应模型的输入数据。

最后是数据存取模块，模块将适应模型的输入数据存入 MySQL 数据库以满足后续工作的数据存取需求。

2.3　新冠疫情数据集的获取

2.3.1　新冠疫情数据集总体情况说明

为了满足后续模型训练的需要，本文利用设计的数据爬取预处理系统爬取了必要的基本数据集。首先是中国的疫情情况数据，包括每天各省市的新增确诊病例数等，时间范围确定为 2020 年 2 月疫情爆发初期一直到 2022 年 3 月，来源首选

国家以及地方卫生健康局官网的疫情通报。其次，在时空点过程模型分析中，本文的主要研究对象是市区，需要得到中国各级行政区划的面积等信息。最后，本文还讨论了深圳市新冠肺炎确诊患者更详细的情况，包括性别、年龄等。因此至少需要收集三个基本数据集，表 2-1 展示了这三个数据集的基本情况：

表 2-1　新冠疫情数据集总体情况说明

数据集名称	主键	数据集大小(单位：条)	主要字段
中国各省市疫情数据集	日期、城市行政代码	381, 156	新增确诊、新增治愈、累计确诊、累计治愈等
中国行政区划数据集	城市（乡镇）行政代码	3, 071	人口数、占地面积、政府所在地名称、政府所在地经度、政府所在地纬度等
深圳市病例数据集	病例编号	1, 084	确诊时间、性别、年龄、常居地名称、常居地经度、常居地纬度

表 2-1 中，中国各省市疫情数据集数据量最大，包含了自 2020 年 2 月疫情爆发初期到 2022 年 5 月精确到二级或三级行政区划（例如：广东省-深圳市属于二级行政区划，北京市朝阳区属于三级行政区划）的疫情数据共约 38 万条。中国行政区划数据集则包含中国各级行政区划的信息情况，包括后续模型训练中需要考虑到的区域面积信息。深圳市病例数据集则是一份较小的数据集，数据条目以患病个体个体为描述对象，主要包含确诊时间和地点信息等。

2.3.2　省市级疫情实时数据的爬取

本部分的数据采集主要依赖于疫情实时播报平台，此时网页更像是一个静态的容器，装入的数据是实时更新的，无法按照爬取网页然后利用 HTML 解析器来解析静态网页的方式获取数据。因此，需要考虑一种动态的方式，从请求网页资源开始，监视服务器端回复的所有格式的文件和数据包，通常数据存放在 json 格式的文件中。通过浏览器的开发者工具，抓取目标 json 数据包的 api 地址，获取和分

析更加格式化的 json 数据。

经过数据对比分析，本课题选择网易的实时发布平台作为爬取对象，将其中有关中国各省市级的病例 json 数据爬取下来，目前累计收集清洗数据量约为 38 万条。如图 2-2 所示，为香港 2022 年 3 月 13 日的疫情情况。可以看到，3 月 13 日香港单日新增确诊 8163 例，需要得到严格的管控，另外毗邻香港的深圳市单日新增确诊 60 例，形势依然严峻，从更大的时空范围看，疫情的爆发的确具有时间和空间的局部性，而基于疫情数据的时空点过程模型的研究，正是把这一种局部性用更形式化的语言加以表征。

图 2-2　数据爬取预处理系统

经过预处理后，将数据存储到 MySQL 数据库中，主要字段包括 LastUpdate（更新日期）、CityShortName（城市名称）、incrConfirmed（新增病例数）、 incrHealed（新增治愈数）incrDeath（新增死亡数）、totalConfirmed（累计确诊数）、totalHealed（累计治愈数）、totalDeath（累计死亡数）等。其中 2022 年 3 月 13 日的部分数据样例如表 2-2 所示：

表 2-2　2022 年 3 月 13 日的疫情情况（部分）

更新日期	城市名称	新增病例数	新增治愈数	新增死亡数
2022-03-13	长春	831	0	0
2022-03-13	青岛	179	0	0
2022-03-13	深圳	60	0	0
2022-03-13	宝鸡	33	0	0

整个省市级新冠疫情数据集包含了自 2020 年 02 月 25 日至今（2022 年 03 月 31 日），同时爬虫部署为每天下午 16:00 执行一次，因此数据集仍在扩充。本文针

对 2020 年 02 月 25 日至 2022 年 03 月 31 日共 765 日约 32 万条数据的数据集进行研究和分析。

2.3.3 深圳市卫健委病例数据的爬取

为了进一步对深圳市新冠肺炎确诊患者数据进行分析，包括本文后续将要开展的基于患者的时空点过程模型分析的研究，本文收集整理了深圳市本土确诊患者数据集。这一数据集粒度更细，精确到每一位患者，需要得到每位患者的详细信息，因此数据预处理过程更加复杂。

首先从深圳市人民政府疫情通报发布页爬取疫情通报，主要包括文章发布时间、文章标题以及文章内容，其中文章内容以字符串列表的形式存放，按照 HTML 中正文内容安排依次存放其中每一个文本标签（p 标签）中的文字，这样处理有利于后续病例信息的提取。

接下来是将每一个本土确诊病例的详细信息从通报中分离出来，主要包含确诊日期、病患性别、病患年龄以及常居地点。这一步首先依据是否含有类似“本土确诊”等字样过滤出含有本土确诊病例详细信息的通报，然后再过滤得到病患的详细信息。最后一步是要把非结构化的常居地转为可处理的结构化数据，也即经纬度，这里利用了百度地图 API，通过查询得到每个地点的经纬度。如表 2-3 所示为深圳市新冠本土确诊患者的详细情况：

表 2-3 深圳市新冠疫情患者的具体信息（部分）

确诊日期	性别	年龄	经度	纬度
2022-02-27	1	49	114.03911	22.56227
2022-02-27	0	9	114.06407	22.53181
2022-02-27	0	37	114.06407	22.53181
2022-02-27	1	44	113.93721	22.50655

数据集合计约有 1000 条记录，每一条记录都代表了一个确诊患者，记录中包含的字段包括确诊日期、病例性别（0 表示女性、1 表示男性）、病例年龄以及常居地经纬度。确诊患者的性别比例为女性: 男性 $=0.9056:1$，可以看到患者中男女比例大致相当，但男性占比稍高。

2.3.4 中国各级行政区划信息的爬取

在时空点过程建模分析的研究中，需要考虑到各个省市的面积这一影响因素，本文收集整理了中国各省、各地级市最新的行政区划信息，例如面积、人口、行政区划代码等等，这些信息可以在中华人民共和国民政部全国行政区划信息查询平台中找到。为了获取到所有地区的信息，本文首先查询并存储了包含 34 个省级行政区划的网页（也就是无法自动翻页），随后将离线页面经过页面解析和过滤预处理得到最终的数据结果。其中北京市的数据如表 2-4 所示：

表 2-4　中国行政区划信息一览（部分，仅北京市）

行政区划代码	地区	面积 单位：km^2	人口 单位：万人	政府部门所在地	邮政编码
110000	北京市	16, 418	1, 392	通州区	（空）
110101	东城区	42	99	景山街道	100010
110102	西城区	51	150	金融街街道	100032
110105	朝阳区	465	214	朝外街道	100020
110106	丰台区	306	116	丰台街道	100071
110107	石景山区	84	39	鲁谷街道	100043
110108	海淀区	431	239	海淀街道	100089
110109	门头沟区	1, 448	25	大峪街道	102300
110111	房山区	1, 995	84	拱辰街道	102400
110112	通州区	906	80	北苑街道	101100
110113	顺义区	1, 020	65	双丰街道	101300
110114	昌平区	1, 342	65	城北街道	102200
110115	大兴区	1, 036	73	兴丰街道	102600
110116	怀柔区	2, 123	29	龙山街道	101400
110117	平谷区	948	41	滨河街道	101200
110118	密云区	2, 226	44	鼓楼街道	101500
110119	延庆区	1, 995	29	儒林街道	102100

表 2-4 给出了北京市及其下级行政单位的基本信息，包括行政区划代码、对应行政区名称以及面积、人口等信息，可以看到北京市一共有 15 个下级行政区划，

占地面积 16, 418km^2，人口共约 1, 392 万人，人口密度平均约为 848 人/km^2。

2.4 新冠疫情数据可视化系统

2.4.1 中国实时疫情信息的可视化分析

为了更直观的对当日的疫情情况作一个直观的了解，本文实现了一个简单的中国疫情信息实时可视化平台。如图 2-3 所示，疫情数据可视化平台主要由中国疫情实时地图、中国疫情的累计趋势和新增趋势折线统计图和新增确诊较多的城市横向条形图构成。

界面左边是两张折线图，左上方是近一周全国累计趋势，其中的三条折线从上至下依次是累计确诊、累计治愈、累计死亡的趋势，左下方则为近一周全国新增趋势，其中的三条折线从上至下依次是新增确诊、新增治愈、新增死亡的趋势，可以看出，无论是累计趋势还是新增趋势，确诊所在折线均远高于其它折线，对比来看，累计确诊数字和新增确诊数字均有较大增幅，形式不容乐观。

区域中部上方精确展示了截至 05 月 16 日全国累计确诊、累计治愈、新增确诊、新增治愈的数字，中部中下方用一张中国地图展示中国各个省级行政区划实时的新增病例情况，新增病例数越多颜色越深。右边则为横向的直方图，从上到下分别表示新增确诊数、新增治愈数最多的十个省市，按照数量递减的顺序进行排列。

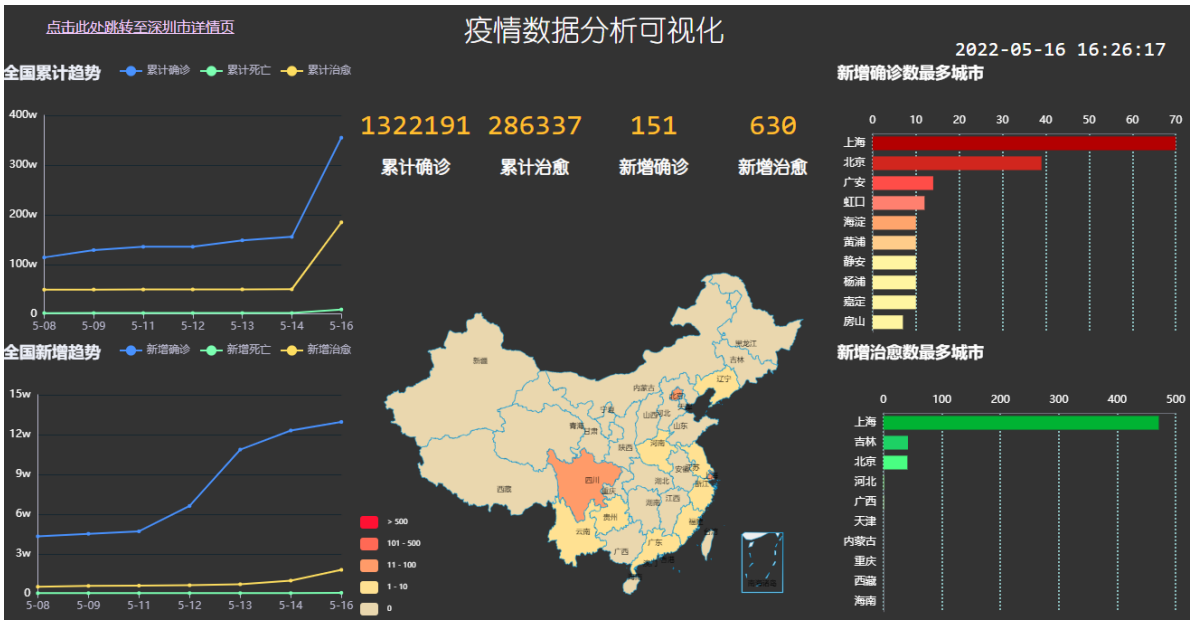


图 2-3 中国疫情实时播报平台概览图（2022-05-16）

这些图表的展示主要运用到了 ECharts 技术。ECharts 的全称是 Enterprise Charts，是一个开源可视化库，提供了丰富的可视化类型，同时兼容大多数数据格式，图表

渲染成本小，具备可交互的特点，易于定制。

本项目首先引入 ECharts 库，以疫情地图地实现为例，需要将库中的地图组件插入到前端的 HTML 文档中，接着从后台查询得到当日的疫情数据，，需要每个省的新增确诊病例数，并划分严重等级，在前台用由浅到深的颜色进行渲染和绘制得到最终的疫情实时地图如图 2-4 所示：

图 2-4　中国疫情新增病例实时地图（2022-05-16）

图 2-4 展示了 2022 年 5 月 16 日全国各省地新冠疫情新增确诊情况，大部分地区处于零新增状态，出现新增病例的省份大多数位于沿海位置，且新增病例数大部分为个位数，其中较严重的是四川省和北京市等地区，疫情的动态清零工作仍需继续。

2.4.2　深圳市疫情信息的可视化分析

针对深圳市新冠病患详细信息数据集，本文也进行了一定的可视化分析工作。如图 2-5 展示了深圳市新冠肺炎患者的年龄分布情况，其中年龄划分以国际常用标准将人的年龄划分为婴幼儿（0~2 岁）、儿童（3~10 岁）、少年（11~17 岁）、青壮年（18~35 岁）、中年（36~60 岁）和老年（大于 60 岁）六个阶段。

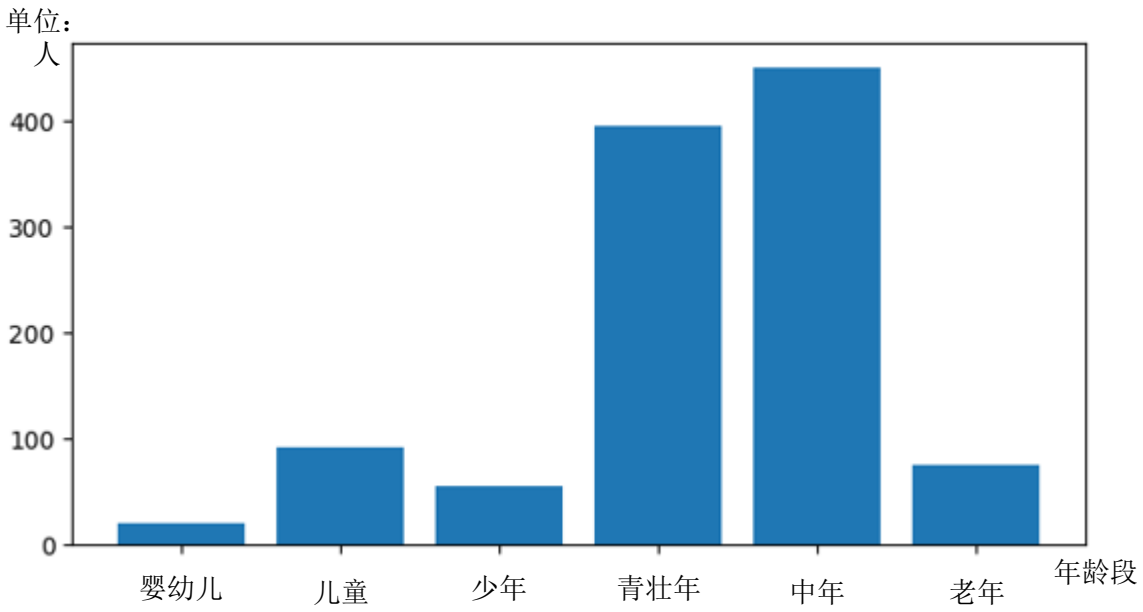
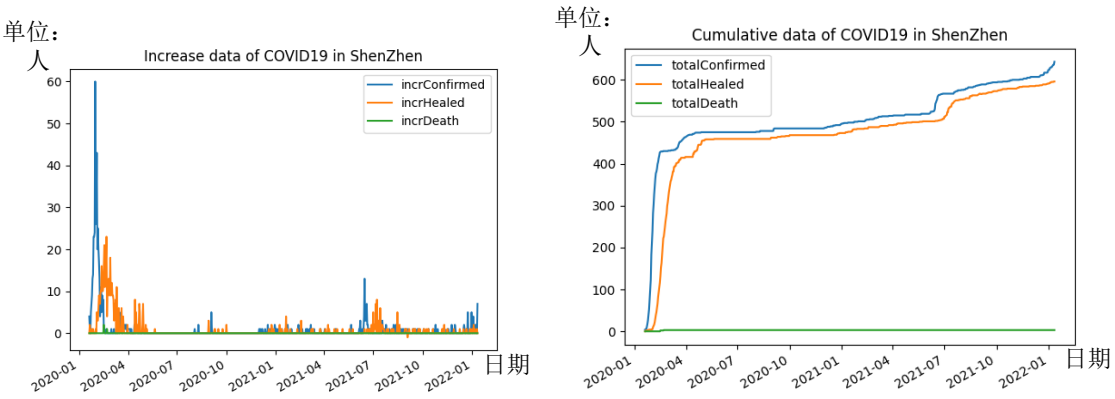


图 2-5　深圳市新冠患者年龄分布（截至 2022 年 03 月 31 日）

图 2-5 表明，在各年龄段中，患病占比最多的是青壮年和中年两个年龄段，尽管年龄划分不均衡，该图仍能说明新冠疫情患病与个体年龄存在一定的联系。一方面观察到青壮年时期为 17 年，从婴幼儿一直到少年阶段结束时期也为 17 年，而青壮年占比要高于婴幼儿一直到少年阶段占比的和，同时中年时期有 24 年，但其占比只是略微高于青壮年的占比，因此总体来看，青壮年更易患病。这一年龄段的人群一般处于就学或就业的环境，流通性和聚集性较强，是导致这一现象出现的可能因素之一。

如图 2-6 展示了深圳市新冠疫情（2020-02 至 2022-03）新冠疫情趋势，其中图(a)为深圳市疫情新增趋势，图(b)为深圳市疫情累计趋势。



（a） 深圳市新冠疫情新增趋势　　（b） 深圳市新冠疫情累计趋势

图 2-6 深圳市新冠疫情（2020-02 至 2022-02）新冠疫情趋势

可以看到深圳市有三轮较大的疫情爆发，分别是疫情扩散初期也即 2020 年 1 月到 4 月、2021 年 7 月份以及 2022 年 1 月底这三次。

2.5 本章小结

本章主要说明了本文使用到的数据来源以及获取方法，介绍了获取网络数据的基本流程，设计实现了数据爬取预处理系统，其中最关键的模块是数据爬取模块，其中的主要程序被称为爬虫。一个高性能爬虫一般具备易扩展、有限制、分布式和可定制四个特点，对于本文数据量相对较少的场景，爬虫只需具备有限制和可定制的特点。同时数据预处理也是一个重要的环节，对于爬取的原始数据，需要依据实验需求过滤筛选、加工转换以得到合适的数据输入。

另一方面，要对爬取到的每日省市数据进行直观地分析，一种好的方法就是将数据进行可视化，并以较丰富的形式呈现，例如折线图、条形图、地图等。

第 3 章　新冠疫情数据的预测分析方法

3.1　引言

在做好充分的文献调研以及数据准备后，需要进一步对疫情数据进行建模分析，主要是对包括 SEIR 模型、改进的 SEIR 模型和 LSTM 模型在内的各类模型进行验证，分析预测疫情的发展趋势并比较各模型的性能优劣。

SEIR 模型以及改进的 SEIR 模型都属于传统的流行病动力学建模中的单一人群模型，这一类模型还包括 SI 模型、SIR 模型等。SI 模型没有设置康复者（R）仓室，适用于感染者不可康复的情况，而 SIR 模型则没有考虑潜伏者（E）的存在，这两种模型均与新冠疫情实际医疗实践情况不符，因而本章着重讨论 SEIR 系列模型。另一方面，本章从时序序列分析的角度采取 LSTM 网络模型进行了模型验证，并对以上三种模型进行了比较。

3.2　SEIR 系列模型原理详述

单一人群模型将整个人群作为研究对象，将人群划分为不同的仓室，通过研究各仓室的人群数量变化规律来预测疫情的发展趋势，其中 SEIR 模型是最经典的传染病模型，包含 S、E、I、R 四类仓室[15]：

(a) **易感者(S)**，不具该病毒免疫力的健康人，易于感染。

(b) **潜伏者(E)**，接触过该病毒但不表现症状的人，可能具备感染性。

(c) **感染者(I)**，感染该病毒并表现出病症的人。

(d) **康复者(R)**，具备该病毒免疫力的人，此后不会再次感染。

四类人群之间的转化关系如图 3-1 所示：

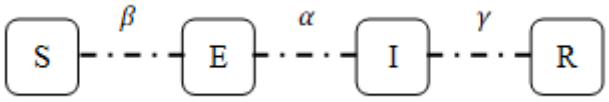


图 3-1　SEIR 模型传播动力图

图 3-1 中包含四类人群和三类转化关系，是从易感者 S 到康复者 R 的单向转化过程。其中β为感染者 I 将病毒传染给易感者 S 的概率，α为潜伏者 E 转化为感染者 I 的概率，γ为感染者 I 转化为康复者 R 的概率。方程组为 SEIR 的动力学方程：

$$\begin{cases} \dfrac{dS}{dt} = -\dfrac{\beta SI}{N} & \text{式(3-1)} \\ \dfrac{dE}{dt} = \dfrac{\beta SI}{N} - \alpha E & \text{式(3-2)} \\ \dfrac{dI}{dt} = \alpha E - \gamma I & \text{式(3-3)} \\ \dfrac{dR}{dt} = \gamma I & \text{式(3-4)} \end{cases}$$

方程组（1）

方程组（1）是一个微分方程组，其中模型假设人群总数固定不变为 N，即有 S+E+I+R=N，同时注意到$\frac{dS}{dt}+\frac{dE}{dt}+\frac{dI}{dt}+\frac{dR}{dt}=0$，这也说明人群的总数不会变化。另外观察到式（3-1）恒为负值，这表现为易感者 S 的数量会持续下降，式（3-4）的值恒为正值，这表现为康复者 R 的数量会不断上升。SEIR 的动力学方程组（1）表明传染病的流行过程是一个单向的过程，从每一个方程式等号右边的算式可以看出，E 是 S 和 I 的中间状态，S 减少的部分（被感染的部分）转移到 E，E 减少的部分（感染后确诊的部分）转移到 I，I 减少的部分（恢复健康并具有免疫力的部分）最后转移到 R。设定初值 N=1000，β=0.272，α=0.0715，γ=0.091，借助数值积分得到 SEIR 方程组的数值解如图 3-2 所示：

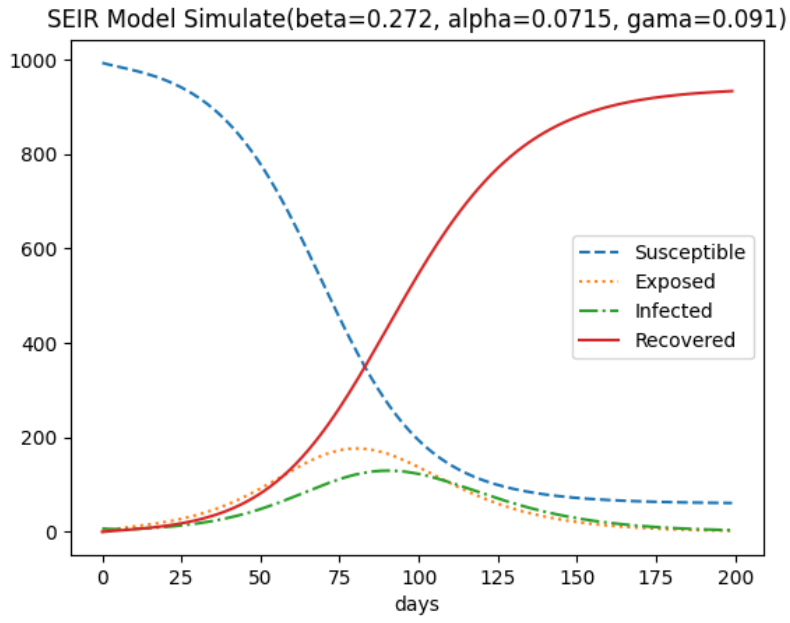


图 3-2　SEIR 模型传播动力图

图 3-2 中横轴表示距疫情爆发以来的天数，纵轴表示四类人群 S（易感者）、E（潜伏者）、I（感染者）、R（康复者）的人数。可以看出，易感者人数持续降低，

康复者人数持续升高，两者均在 150 天左右到达稳定状态。而潜伏者和感染者的人数走势类似，都是先升高达到峰值后回落，区别在于：潜伏者峰值较高且出现时刻更早。

针对新冠疫情的实际情况，首先人群中存在潜伏者 E，其次医疗实践表明潜伏者 E 也具备一定的感染性，据此对方程组（1）作出修正得到方程组（2）：

$$\begin{cases} \dfrac{dS}{dt} = -\dfrac{\beta_1 SE + \beta_2 SI}{N} \\ \dfrac{dE}{dt} = \dfrac{\beta_1 SE + \beta_2 SI}{N} - \alpha E \\ \dfrac{dI}{dt} = \alpha E - \gamma I \\ \dfrac{dR}{dt} = \gamma I \end{cases}$$

方程组（2）

相对于方程组（1），方程组（2）中将参数β替换为β_1和β_2，其中β_1参数表征了潜伏者 E 对易感者 I 的感染率。修正后的 SEIR 模型各类人群转化关系如图 3-3 所示：

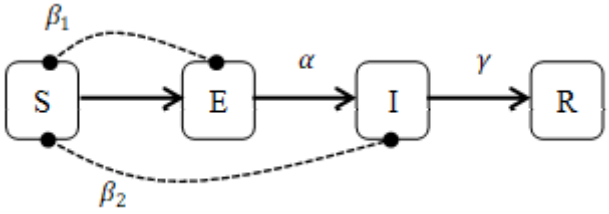


图 3-3　修正 SEIR 模型的传播动力图

在修正 SEIR 模型的传播动力图中，四类人群仍然是单向转化，不同之处在于易感者可以因接触潜伏者而有β_1的几率转化为潜伏者。在方程组（2）中设定初值 N=1000，β_1=0.127，　β_2=0.145，α=0.0715，γ=0.091，借助数值积分得到修正的 SEIR 方程组的数值解如图 3-4 所示：

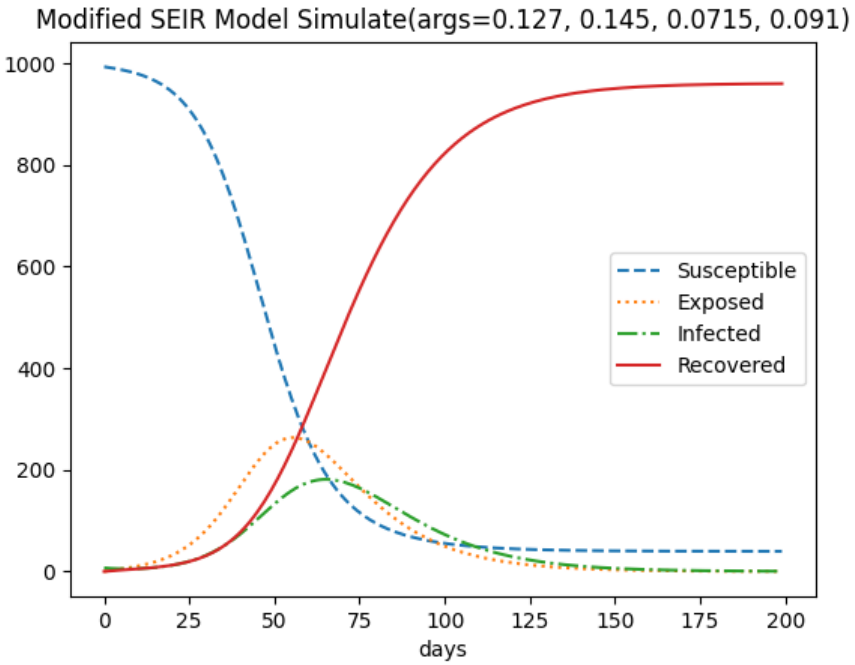


图 3-4　修正 SEIR 模型传播动力图

相对于 SEIR 模型，考虑到潜伏者也会感染易感者的修正 SEIR 模型（以下简称 mSEIR 模型）其各类人群数量的变化趋势并没有变化，但此时潜伏者、感染者峰值数量均有所升高，峰值出现时间也有所提前，较好地符合了修正后的假设。

3.3　基于深圳市数据的 SEIR 系列模型分析

SEIR 系列模型假设人群的每个个体相同，人群均匀混合（homogeneous mixing）；接触是瞬时的，接触与历史无关；每个仓室的人群数量足够大，在传染病流行过程中，感染率、恢复率为常数。因此这一类方法一般适用于疫情爆发初期短时间内某个城市的疫情趋势（走向）的预测和判断，本文以深圳市的疫情数据为例展开对 SEIR 系列模型的验证分析。

首先，需要从中国各省市疫情情况数据集中查询得到深圳市 2020 年 02 月到 2022 年 02 月的疫情数据，包括每日的新增\累计本土确诊人数、新增\累计死亡人数、新增\累计治愈人数，截取第一次爆发的数据来对 SEIR 模型、mSEIR 模型的参数进行拟合，时间范围从 2020 年 01 月 20 日到 2020 年 04 月 15 日。同时，对于实际的易感者、潜伏者、感染者和康复者数量，可以确定的是两个，康复者数量也就是累计治愈数（Heal），而感染者数量用累计确诊数（Confirm）减去累计治愈

数再减去累计死亡数（Dead）得到。计算公式如式(3-5)、式(3-6)所示：

$$I(t) = Confirm(t) - Heal(t) - Dead(t) \quad \text{式(3-5)}$$

$$R(t) = Heal(t) \quad \text{式(3-6)}$$

同时，潜伏者数量 E(t)一般可以用无症状感染者数量代替，但由于早期检测手段和技术还不够成熟，又由于潜伏者本身就意味着没有症状，难以发现，检测代价较大，因此没有相应的数据。然而，如果从实际疫情防控需要的角度出发，对于$I(t)$峰值点的预测是其核心，包括何时会到达疫情传播高峰期以及此时感染者的数量，对于其它三类人群的关心程度更少。

对于方程参数确定问题，首先通过参数的实际意义结合医疗实践对部分参数进行，包括α感染率和γ治愈率，依据隔离 14 天的标准取α=1/7, γ=1/14，β定义一个合理初值例如 0.08。使用均方误差函数和梯度下降算法对α和γ和β进行调整，最后得到α=0.2, β=0.0035, γ=0.04。最终的拟合图像如图 3-5 所示：

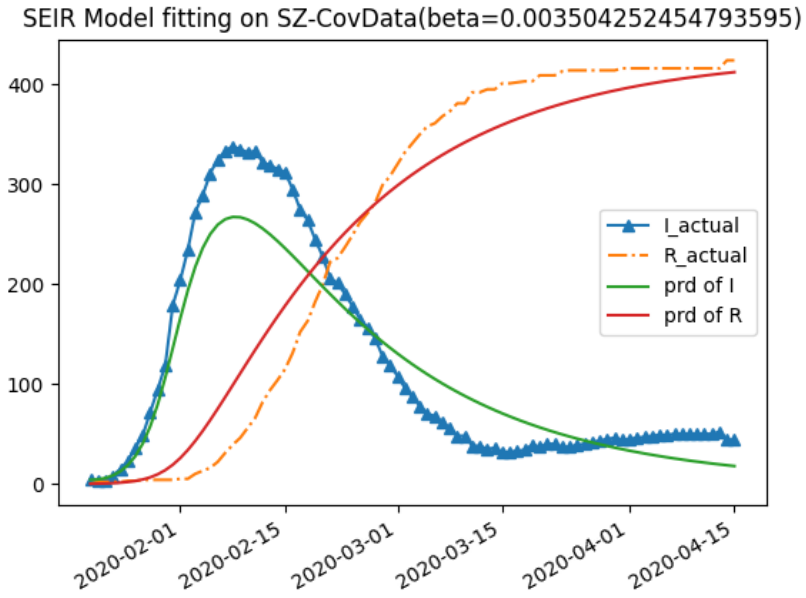


图 3-5　深圳市新冠疫情 SEIR 模型的参数拟合

图 3-5 中两条带点的曲线为实际数据，另外两条不带点的光滑曲线为拟合曲线，可以看到对于感染者数据预测效果较好，峰值出现日期基本一致，在 2 月 8 日左右；另一方面峰值点高度预测值低于实际值，出现这一现象的原因在于实际情况与 SEIR 模型假设存在一定程度的不符，例如，这一时期内会有境外输入导致感染率升高，与感染率恒定的假设不符；疫情爆发初期人们需要一定的反应时间，医药方法的研究还未及时跟进，治愈率相对较低等。

针对 mSEIR 模型，由于使用的数据集相同，且α和γ具备实际意义，从而可

以将这两个参数固定下来，对β_1和β_2进行参数拟合，同样使用均方误差作为误差函数，使用梯度下降算法进行拟合。结果如图 3-6 所示：

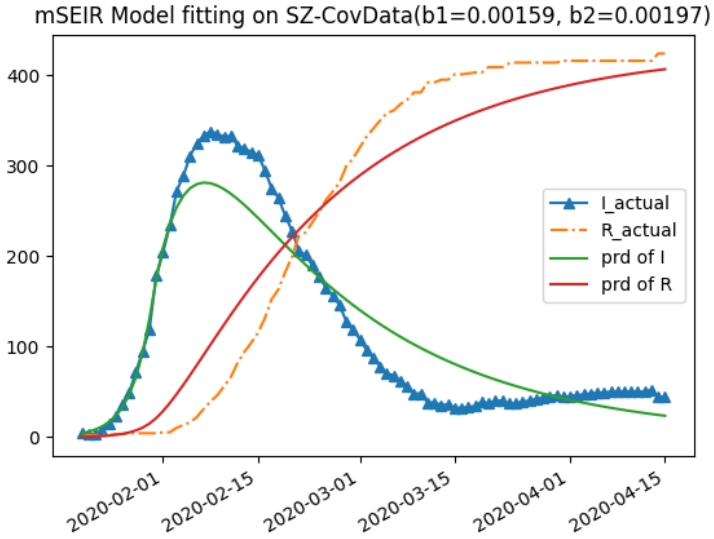


图 3-6　深圳市新冠疫情 mSEIR 模型的参数拟合

在改进的 SEIR 模型参数拟合图像中，感染者预测曲线在到达峰值前与实际值较原 SEIR 模型更加吻合。两个模型都能较好的预测峰值来临的日期，而峰值点人数预测值均偏低，在峰值点来临后，相较于实际患病人数曲线，预测曲线下降更缓慢；另一方面观察康复者曲线，预测曲线与实际曲线均呈现单调上升的状态，两曲线之间有一个交点，在相交之前，预测值偏高，而在相交之后，预测值又偏低，随着时间推移，两曲线又再次接近并趋向稳定。

从上述描述和分析可以得出：首先 SEIR 系列模型能够在一定程度上预测实际新冠疫情初期的发展趋势，能够作为实际医疗实践和人工干预措施的科学参考。其次由于模型较为严苛的假设，预测曲线与实际曲线又存在偏差，一个合理的猜想是感染率和治愈率并不是一直为常数，在疫情爆发起始时，由于人工干预还未到位以及对新冠病毒的医疗实践不够，新冠疫情感染率较高而治愈率较低，导致初始时感染者不断累积滞留，很快到达一个较高的峰值，此时康复者数量较少，上升缓慢；此后随着人工干预措施例如隔离等的实施，使得感染率大幅降低，感染者累积速度大大放缓，另一方面康复率变高，两者结合导致感染者数量下降较为迅速，康复者数量上升也较快，最后随着感染者的持续减少，康复者的持续增多直到趋于稳定，表明一轮疫情传播的结束。

3.4　基于长短期记忆网络的时序分析

3.4.1 RNN 网络模型的工作原理和存在的问题

递归神经网络（Recurrent Neural Networks，RNN）是一种可以处理顺序数据或者是时间序列数据的人工神经网络，是神经网络的一般类别，已经在语音识别、时间序列预测和自然语言处理等领域得到广泛的应用。

在递归神经网络中，模型输出依赖于有序输入的历史序列。例如向递归神经网络输入一个句子，句子中的单词是按照一定顺序排列的，“他在那里”和“在他那里”两个句子语义是完全不一样的，当递归神经网络得到“他”和“在”两个单词，下一个单词可能是正在进行的活动，也可能是某个地区区域等，为了判断哪一种情况的可能性最大，往往需要向前看更多的历史输入，但当历史事件间隔过长时，可能无法从中获得有效信息，也即产生了长期依赖问题[3]。

递归神经网络的另一个特征是它们在网络的每一层共享相同的参数（权重），在反向传播和梯度下降的过程中统一调整以促进学习。因而，在递归神经网络的模型训练过程中，例如进在行股票市场预测、机器翻译和文本生成等工作训练的过程中，梯度消失问题以及梯度爆炸问题使得递归神经网络变得难以训练，当目标子序列中的依赖关系跨越大量样本时，要求递归神经网络的工作窗口相当宽，此时这些困难变得明显[3]。

3.4.2 LSTM 网络的工作原理与改进点

长短期记忆网络（Long-Short Term Memory, LSTM）是一种特殊的 RNN，针对原始的 RNN 结构存在的问题，LSTM 对循环节点结构做出了改进，相对于 RNN 更加复杂，如图 3-7 所示为 LSTM 的网络结构[4]：

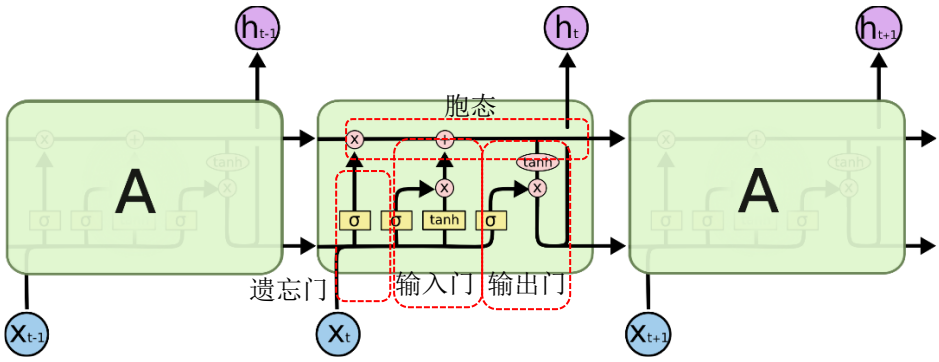


图 3-7　LSTM 网络模型结构

LSTM 网络模型的输出结果依赖于当前时刻的细胞状态（$cell\ state$）、前一时刻的隐含层输出（h_{t-1}）以及当前时刻的数据输入（x_t）。图 3-8 展示的 LSTM 网

络的循环节点是产生输出结果的核心部分，节点划分为四个部分或称四个门控，相比于原始递归神经网络要复杂不少，门控机制的引入在很大程度上缓解了长期依赖问题和梯度问题[8]，是对递归神经网络的极大改进。

从 LSTM 网络产生输出结果的过程来阐释：首先是遗忘门（$forget\ gate$），遗忘门依据上一时刻的隐含层输出（h_{t-1}）以及当前时刻的数据输入（x_t）的$sigmoid$值（f_t）来决定上一时刻细胞状态（C_{t-1}）的哪些位可以传入到当前的细胞状态（C_t），总体来看，遗忘门决定了上一时刻的状态中哪些需要被去掉。其中f_t的计算方法如式(3-7)所示（计算值落在区间[0, 1]内）：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \qquad \text{式(3-7)}$$

其次是输入门（$input\ gate$），输入门本身就是一个神经网络。它首先经过$tanh$层生成新的记忆向量，指明当前状态的更新部分，然后通过$sigmoid$激活函数（过滤器）识别哪些更新的部分是认为有效的，如式(3-8)、式(3-9)、式(3-10)所示，再结合遗忘门输出即可更新细胞状态。

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \qquad \text{式(3-8)}$$

$$\overline{C_t} = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \qquad \text{式(3-9)}$$

$$C_t = f_t * C_{t-1} + i_t * \overline{C_t} \qquad \text{式(3-10)}$$

如式(3-11)和式(3-12)所示，通过$sigmoid$激活函数得到模型的输出O_t，结合已更新的细胞状态得到当前的隐含层输出h_t。

$$O_t = \sigma(W_O \cdot [h_{t-1}, x_t] + b_O) \qquad \text{式(3-11)}$$

$$h_t = O_t * \tanh(C_t) \qquad \text{式(3-12)}$$

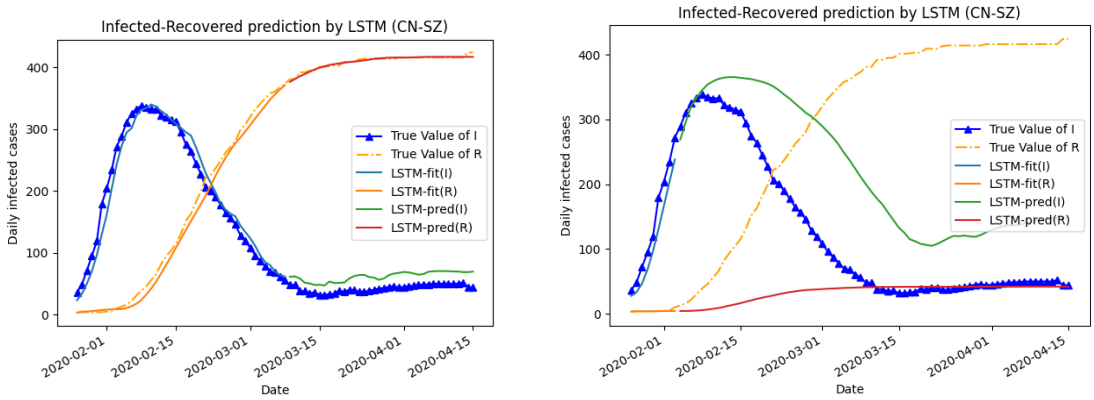
3.4.3 使用 LSTM 网络分析疫情数据

LSTM 网络可以借助 python 的依赖库 pytorch 中的 nn.Module 类进行实现，整个过程分为训练和预测两个部分，训练时采用 MSE 作为误差函数，MSE 即 Mean Square Loss，表示平方平均误差，基本形式如式(3-13)所示：

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - y_i)^2 \qquad \text{式(3-13)}$$

在进行模型训练时，首先定义训练比（$ratio$），也即用于训练的数据占比，考虑两种情况：一是使用一个通常情况下较高的训练比，例如 0.60；二是针对实际新冠疫情数据预测而言，如果只将峰值之前的数据用于训练，模型能否预测到峰值点的到来，此时训练比为 0.11。

在两种情况下，LSTM 模型训练结果如图 3-9 所示：



（a） 训练比=0.60 时 LSTM 的工作情况　　（b） 训练比=0.11 时 LSTM 的工作情况

图 3-8　不同训练比情况下 LSTM 的工作情况

图 3-8 表明，在有充分的训练数据时，LSTM 网络能够很好的学习到实际疫情发展的趋势，预测值与实际值十分吻合；而在训练样本容量较小时，LSTM 模型完全没有学习到康复者数量变化的趋势，但感染者的总体变化趋势预测是正确的，LSTM 网络预测到了峰值的出现，而不是持续的单调上升或者放平。因此，当得到充分的训练样本数据时，LSTM 网络模型的表现较好。

3.5　模型间的比较与分析

本文使用了统一的误差评判标准 RMSE（均方根误差）和 MAE（平均绝对误差）对几种模型的预测结果进行了计算和评判，结果如表 3-1 所示：

表 3-1　各类模型在深圳市早期新冠疫情数据集上的预测情况

模型	RMSE	MAE
SEIR	34.48688	28.051678
mSEIR	34.33434	27.573588
LSTM(ratio=0.60)	15.64497	12.75899
LSTM(ratio=0.11)	196.01667	169.45608

整体来看，LSTM（ratio=0.60）模型预测值与实际值最为接近，在训练数据足

够时，LSTM 模型能够学习到实际疫情的发展趋势，另一方面 LSTM（ratio=0.11）模型预测值与实际值偏差最大，在训练数据不足时，LSTM 模型不足以学习到实际疫情的发展趋势，这说明了 LSTM 网络模型是不依靠先验知识或预先假设的，而是依赖于数据并从中进行学习。而 SEIR 模型和 mSEIR 模型总体误差接近，都表现较好，而 mSEIR 模型略胜一筹，这两个模型依赖于数学假设，与实际情况难免有偏差，但另一方面其内在规律又大体能够描述疫情发展趋势，因此能够较好地预测到感染者峰值点到来的日期。

3.6 本章小结

本章主要讨论了 SEIR 系列模型和 LSTM 模型的基本假设与设计原理，预置 SEIR 系列模型的参数进行了仿真，值得注意的地方是感染者数量曲线，它是先上升后下降的，从而有一个峰值点，这是本章研究的重点。同时，本章还阐述了 LSTM 模型的相对于原始 RNN 网络的几点改进之处，并对 SEIR 模型的假设提出了修正，主要是加入了“潜伏者有几率接触感染易感者”的因素，同时对方程也做了相应的修正。

最后，为了进一步探讨模型的实际性能，本章还抽取了深圳市疫情爆发初期的数据，对以上三类模型进行了训练和参数拟合，并将结果绘制为图表的形式，评价了他们的优劣与适用性。

第 4 章　基于时空点过程的新冠疫情数据分析

4.1　引言

在第 3 章中，本文主要针对疫情发展趋势的分析和预测这一问题分别从传统的单一人群仓室模型和近年来在深度学习中应用广泛的时序序列分析模型 LSTM 网络两个角度进行了研究和讨论。另一方面，由于模型假设的限制，在第 3 章中使用到的数据集是中国各省市历史疫情数据的子集，包括空间范围限制在深圳市而时间范围限制在 2020 年 2 月第一轮疫情爆发到 4 月扩散基本结束，因而数据利用不够充分。

本章采用最新流行的时空点过程（Spatial-Temporal Point Process）视角对中国各省市历史疫情数据集进行了分析和研究，时空点过程分析方法的研究对象是一系列分布在连续时空域的抽象点（或称事件）[8]，每一个事件由一个时空坐标唯一确定，事件也可以有附加属性。顾名思义，时空点过程分析方法下有两个子研究分支，包括时间点过程分析法和空间点过程分析法，本章主要围绕 Poisson 点过程、Hawkes 点过程、神经 Hawkes 点过程等模型展开对中国的疫情数据的研究分析。

4.2　时空点过程的事件建模法

4.2.1　时空点过程的一般形式

令$H = \{t_i,\ \boldsymbol{x_i}\}_{i=1}^{n}$表示一个事件序列，其中$t_i \in R$，$\boldsymbol{x_i} \in \boldsymbol{R^d}$是对应的空间坐标，同时，令$H_t = \{(t_i,\ \boldsymbol{x_i})|t_i < t,\ t_i \in H\}$表示发生在$t$时刻之前的所有事件。如式(4-1)所示，一个时空点过程可以使用一个条件概率强度函数完整刻画[13]：

$$\lambda(t,\boldsymbol{x}|H_t) \triangleq \lim_{\triangle \boldsymbol{x}\to 0,\triangle t\to 0} \frac{P(t_i \in [t,t+\triangle t],\ \boldsymbol{x_i} \in B(\boldsymbol{x},\triangle \boldsymbol{x})|H_t)}{|B(\boldsymbol{x},\triangle \boldsymbol{x})|\ \triangle t} \quad \text{式(4-1)}$$

式 4-1 中的$B(\boldsymbol{x},\triangle \boldsymbol{x})$一个定义在$R^d$上的以$\boldsymbol{x}$为中心以$\triangle \boldsymbol{x}$为邻域半径的带心领域（相对于去心邻域）。注意到$\lambda(t,\boldsymbol{x}|H_t)$必须是一个正数（概率的实际意义），不妨将其简记为$\lambda^*(t,\boldsymbol{x})$，也即$\lambda^*(t,\boldsymbol{x}) = \lambda(t,\boldsymbol{x}|H_t)$。在时间间隔$[0,T]$内对于观测值$H(t)$的联合对数似然值由式(4-2)给出[16]：

$$\log p(H) = \sum_{i=1}^{n} \log \lambda^*(t_i, \boldsymbol{x_i}) - \int_0^T \int_{\boldsymbol{R^d}} \lambda^*(\mu, \boldsymbol{x}) d\boldsymbol{x} d\mu \qquad \text{式(4-2)}$$

式(4-2)中包含多重积分，因此想要以此为极大似然函数训练一个一般性的STPP（Spatial-Temporal Point Process）模型是十分困难的。为了保证模型仍然能够使用较高精度的似然估计函数，本文采用 RTQ Chen 的方法，利用 Neural ODE（Neural Ordinary Differential Equations）框架对 STPP 模型进行参数化[16]，这样就简化了计算过程。

如式(4-3)所示简化过程的第一步是将$\lambda^*(t, \boldsymbol{x})$进行分解：

$$\lambda^*(t, \boldsymbol{x}) = \lambda^*(t) p^*(\boldsymbol{x}|t) \qquad \text{式(4-3)}$$

式(4-3)中的*符号仍然表示某事件在条件H_t下的概率，那么$\lambda^*(t)$表示相对应的时间点过程的偏概率强度函数，$p^*(\boldsymbol{x}|t)$表示在H_t条件下在时刻 t 存在事件在位置$\boldsymbol{x}$发生的条件概率，因此式(4-2)可以简化为：

$$\log p(H) = Spatial + Temporal \qquad \text{式(4-4)}$$

$$Temporal = \sum_{i=1}^{n} \log \lambda^*(t) - \int_0^T \lambda^*(\mu) d\mu \qquad \text{式(4-5)}$$

$$Spatial = \sum_{i=1}^{n} \log p^*\left(\boldsymbol{x}_{\boldsymbol{t_i}}^{(\boldsymbol{i})} | t_i\right) \qquad \text{式(4-6)}$$

因此，一个时空点过程可以使用一个和函数表示，一部分是时间点过程条件强度函数，另一部分是空间点过程的推荐强度函数，此时整个方程组中只包含一个一重积分（对时间）。

4.2.2 几种点过程模型的探讨

（1） 泊松过程（Poisson Process）

在 4.1.1 中的讨论中可以知道，使用一个条件强度函数$\lambda(t)$可以完全地表示一个点过程。泊松点过程认为$\lambda(t)$是独立于历史事件的，其中同质泊松过程（homogeneous）认为$\lambda(t)$是一个常数，即$\lambda(t) = \lambda_0$；而非同质泊松过程则认为$\lambda(t)$是随时间变化的，即$\lambda(t) = g(t)$。

（2） 自校正过程（Self-Correcting Process）

自校正过程认为强度函数的趋势是一直在增大，但是当一个事件发生后，会先减小，式(4-7)为自校正过程的强度函数：

$$\lambda(t) = e^{\mu t - \sum_{t_i < t} \alpha} (\alpha > 0, \mu > 0) \qquad \text{式(4-7)}$$

自校正过程的强度函数是以欧拉数为底数的指数函数的形式，其中指数部

分的$\mu t(\mu > 0)$表示强度函数随着时间的推移持续的增大，而$-\sum_{t_i<t}\alpha$部分则表明每当有事件发生时，强度函数的值会减小。

（3） 霍克斯点过程（Hawkes Point Process）

霍克斯过程是一种强度函数值依赖于历史事件的自激励点过程，其特征是其历史事件的影响以时间累加的形式进行，霍克斯过程的强度函数定义如式(4-8)所示[21]。其中H_t表示截止到时间t的所有历史事件，φ_0是给定的确定性函数，ϕ是核函数，描述了过去事件对当前事件的影响。也就是说，这是对在时间点t之前的所有事件都进行了考虑。为了描述这些影响随着时间的衰减效应，尤其是最新事件对未来事件的最大影响，核函数一般采用类似于指数函数的形式随着时间变化的强度函数的实例如图 4-1 所示。

$$\lambda(t|H_t) = \varphi_0(t) + \sum_{t_i<t}\phi(t, t_i) \qquad \text{式(4-8)}$$

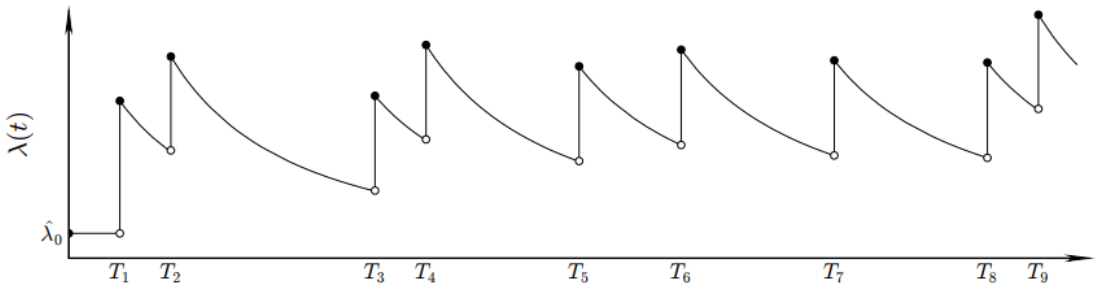


图 4-1　核函数的变化曲线

图 4-1 展示了霍克斯过程的强度函数曲线随时间推移发生的变化，可以看到历史事件对未来事件的影响是单调指数递减，然后以累加形式进行叠加。其中横轴标注了各个事件发生的时间，曲线上的空心圆点则代表着事件，每当有新事件发生时，事件产生一个正向激励，使强度函数呈跳跃式增长。

以上三种点过程属于传统点过程，都具有较强的解释性，需要一定的先验知识，并利用一定的数学假设和公式对条件强度函数进行刻画。然而，传统点过程对强度函数有着上述设定，很有可能不符合实际情况，比如历史事件对强度函数的影响并不一定是累加的；此外，如果有多种事件类型的话，还需作出各个事件类型是互相独立的假设，并且对每个事件类型求强度函数。

深度点过程则只需使用神经网络这样的非线性函数模拟强度函数，这样一个黑盒子无需设定任何先验知识。本文研究的一个深度点过程是基于霍克斯过程的神经霍克斯点过程，采用 Mei H 教授的训练方法[17]利用神经网络模拟了霍克斯过程的核函数。

4.3 基于点过程模型的新冠疫情数据分析

4.3.1 数据的预处理

此处需要用到的数据集中国各省市历史疫情数据的全部有效数据，也即需要过滤出所有包含新增本土确诊病例的记录，每条记录被日期和城市唯一确定，最终过滤得到的数据一共约 8000 条。

其次，利用城市代码在中国行政区划信息表中查询得到面积，利用百度地图 API 获取得到经纬度，以 2020 年 02 月 01 日为起点，将日期转为距起始日期的天数 days 这样就得到了初始输入，如表 4-1 所示：

表 4-1 预处理后得到的初始输入

天数	地区	新增病例数	面积 Km^2	经度	纬度
0	恩施州	105	24, 111	109.49459	30.27794
0	黔南州	1	26, 195	107.52840	26.26062
0	黔东南州	2	30, 339	107.98945	26.58970
0	福州	5	11, 974	119.30347	26.08043
0	泉州	5	11, 015	118.68245	24.87995
0	莆田	2	4 , 132	119.01452	25.45987
0	厦门	1	1, 701	118.09644	24.48541
0	漳州	1	12, 882	117.65358	24.51893
0	三明	1	22, 965	117.64552	26.26974
0	温州	24	11, 784	120.70648	28.00109
0	杭州	12	16, 596	120.21551	30.25308

4.3.2 点过程模型的部署与训练

模型的建立和训练主要依赖于 python 的第三方库 pytorch，并使用 CUDA 工具借助 GPU 进行加速，使用对数似然值对模型的结果进行评价分析。

最大似然估计法的思想在于：在已经得到试验结果的情况下，我们应该寻找使这个结果出现的可能性最大的那个x^*作为真值x的估计，因而问题转化为为求

极大似然函数$L(X)$的最大值点，而由于$L(X)$往往是乘积的形式并且对数函数是单调增函数，取对数后求解更简单。对$\log L(X)$关于X求导数，并命其等于零，得到的方程组称为似然方程组。解方程组得到X的最大似然估计。

结合式 4-5 和式 4-6，通过 python 的 scipy 库中的数值积分包进行计算，得到数据结果如表 4-2 所示：

表 4-2 几种点过程模型在中国疫情数据集上的训练结果

model	Spatial log-likelihood	temporal log-likelihood
Poisson	-3.19491±1.23	0.17915±0.023
self-correcting	-2.19491±1.24	0.25971±0.022
Hawkes	-0.39191±0.012	0.14253±0.023
neural Hawkes	-0.38132±0.022	0.04132±0.013

结果表明，总体来看，几个过程的空间对数似然值差别不大，表现都比较好。在时间对数似然值上，泊松过程模型和自校正模型表现较差，其关于强度函数的假设与实际新冠疫情爆发的时空过程有较大偏差，而基本的霍克斯过程实验结果已经表现相当良好，但基于神经的神经 STPP 可以实现更好的空间可能性，这也说明了神经网络的优势在于不需要或者只需要很少的先验知识和假设，而是通过网络来模拟其条件强度函数。

4.4 本章小结

本章首先介绍了时空点过程的一般形式，指明条件强度函数能够完全地刻画一个时空点过程，给出了时空点过程的一般形式。然而，由于其对数似然函数包含多重积分，想要直接进行训练是困难的，因而可以采取参数化的方法将对数似然函数简化以适合用于训练。其次介绍了几个典型的点过程模型的方法原理，包括泊松点过程模型、自校正点过程模型和霍克斯点过程模型，并基于中国各省市历史疫情数据集对以上三个模型以及神经霍克斯过程模型进行了训练。最后利用对数似然值对模型进行评价分析，相较于其他模型，霍克斯模型以及神经霍克斯模型在中国的历史疫情数据集上有着更为出色的表现。

结　论

本文收集了较为详细的中国各省市的历史疫情数据，同时还包括了中国各行政区的基本信息，形成了一个较为全面且准确的数据库，并针对性的完成了一系列的可视化工作，包括疫情趋势走向折线图、疫情地图等等。

本文首先从传统的仓室模型中选取了 SEIR 模型进行了模拟，同时对从“潜伏者也可能接触并感染易感者”这一角度对 SEIR 模型进行修正后的模型也进行了模拟，结果表明感染者曲线呈现先升后降的趋势，其顶点位置的预测成为工作的重点。随后本文利用深圳市的真实数据集对 SEIR、mSEIR 模型进行了参数拟合，在训练过程中，本文创新性的将感染率、恢复率等参数通过医疗实践知识进行预确定，从而可以舍弃不确定的潜伏者数据，使其余参数的拟合过程依赖于确定的数据。

此外，本文创新性的跳出传统的传染病动力学的分析角度，将疫情数据视为时序数据，并针对 LSTM 模型进行了实验，结果表明在相同的深圳市早期疫情数据集上，当训练样本容量充足时，LSTM 网络模型预测值与真实值最为接近，同时 SEIR 系列模型也能够较好的预测新冠疫情的初期发展趋势。进一步的，将空间坐标也考虑进来，本文创新性的从时空点过程分析方法在中国历史疫情数据集上展开了几个时空点过程模型的训练，并利用对数似然值对模型进行了评价分析。

针对本文的工作，还可以从以下角度进行进一步的研究分析：

（1）SEIR 模型往往只能模拟一轮疫情爆发，且其中在疫情散播过程中不会变，能否借助深度学习模型识别新的一轮疫情爆发，清空旧的参数，在一轮疫情扩散过程中对参数进行自适应的调整；

（2）针对 SEIR 模型本身，还可以做许多其它的修正，例如加入复阳者人群，他们在康复后可能复发；在开发出疫苗后，一部分人群，他们提前接种了疫苗，可以考虑加入预接种人群等；

（3）针对时空点过程，还可以做更深入的分析，例如基于深圳市详细病例数据集的研究，同时需要做更深入的探讨使结果更清晰易懂。

参考文献

［1］ 张发,李璐,宣慧玉. 传染病传播模型综述[J]. 系统工程理论与实践,2011,31(9):1736-1744.

［2］ Kermack W O, McKendrick A G. Contributions to the mathematical theory of epidemics, part I [J]. Proceedings of the Royal Society of London A, 1927, 115:700-721.

［3］ Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network[J]. Physica D: Nonlinear Phenomena, 2020, 404: 132306.

［4］ Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural computation, 2019, 31(7): 1235-1270.

［5］ 赵永翼, 王菲, 申莹. 基于长短期记忆网络的 COVID-19 疫情趋势序列分析预测[J]. 沈阳师范大学学报 (自然科学版), 2020.

［6］ Ogata Y. Space-time point-process models for earthquake occurrences[J]. Annals of the Institute of Statistical Mathematics, 1998, 50(2): 379-402.

［7］ Stoyan D, Penttinen A. Recent applications of point process methods in forestry statistics[J]. Statistical science, 2000: 61-78.

［8］ Junchi Yan. Recent Advance in Temporal Point Process: from Machine Learning Perspective[R]. Think Lab 2019.

［9］ David R Cox. Some statistical methods connected with series of events. Journal of the Royal Statistical Society. Series B (Methodological), pages 129–164,1955.

［10］A. G Hawkes. Spectra of some self-exciting and mutually-exciting point processes. Biometrika,58, 1971

［11］Sebastian Meyer, Johannes Elias, and Michael Hohle. A space–time conditional intensity model for invasive meningococcal disease occurrence. Biometrics, 68(2):607–616, 2012.

［12］Junhyung Park, Adam W Chaffee, Ryan J Harrigan, and Frederic Paik Schoenberg. A non-parametric Hawkes model of the spread of ebola in west africa. 2019.

［13］RTQ Chen, Amos B, Nickel M. Neural spatio-temporal point processes[J]. arXiv preprint arXiv:2011.04583, 2020.

［14］周德懋,李舟军.高性能网络爬虫:研究综述[J].计算机科学,2009,36(08):26-29+53.

［15］甘雨,吴雨,王建勇.新冠肺炎疫情趋势预测模型[J].智能系统学报,2021,16(03):528-536.

［16］Adrian Baddeley, Imre Bar´ any, and Rolf Schneider. Spatial point processes and their applications. ´Stochastic Geometry: Lectures Given at the CIME Summer School Held in Martina Franca, Italy,September 13–18, 2004, pp. 1–75, 2007.

［17］Mei H, Eisner J M. The neural hawkes process: A neurally self-modulating multivariate point process[J]. Advances in neural information processing systems, 2017, 30.

哈尔滨工业大学深圳校区本科生毕业设计（论文）原创性声明

本人郑重声明：在哈尔滨工业大学深圳校区攻读学士学位期间，所提交的毕业设计（论文）《新冠疫情时空数据的自动采集与分析方法》，是本人在导师指导下独立进行研究工作所取得的成果。对本文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明，其它未注明部分不包含他人已发表或撰写过的研究成果，不存在购买、由他人代写、剽窃和伪造数据等作假行为。

本人愿为此声明承担法律责任。

作者签名：段裕　　日期：2022 年 05 月 27 日

致　谢

衷心感谢导师冯山山教授对本人的精心指导，冯山山教授是一个平易近人、善解人意的人，在毕设开题初期，由于本人在准备研究生考试，他让我放下心理负担，专心备考；在考试后，冯老师的关心和指导使我能够踏实的一步步的将毕业设计做好，他的言传身教将使我终生受益。

感谢我的父母，疫情期间无法按时返校，是他们尽力为我创造了相对良好的学习研究环境！

感谢辅导员辛欣老师，辛老师温柔体贴，给予了我关心、鼓励和支持，帮助我走出情绪低落期。

感谢哈尔滨工业大学（深圳）实验与创新实践教育中心的支持，使我能够顺利的完成模型的训练和数据分析。

感谢所有的奋战在一线的抗疫工作者们，你们都是最棒的！没有你们，我们就无法回归到基本正常的轨道，你们辛苦了！