

本科毕业论文（设计）

题 目 基于流感监控数据的新冠肺炎预测方法研究

学 院 计算机与信息科学学院 软件学院

专 业 计算机科学与技术

年 级 2017级

学 号 222017321102067

姓 名 刘炳辰

指 导 教 师 陶丽

成 绩 \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

年 月 日

目 录

[摘 要： 1](#_Toc72341026)

[Abstract: 1](#_Toc72341027)

[第1章 导论 3](#_Toc72341028)

[1.1研究背景 3](#_Toc72341029)

[1.2主要工作 4](#_Toc72341030)

[1.3本文结构 5](#_Toc72341031)

[1.4本章小结 6](#_Toc72341032)

[第2章 相关研究和技术方法概述 7](#_Toc72341033)

[2.1相关研究 7](#_Toc72341034)

[2.2相关技术 8](#_Toc72341035)

[2.2.1计算疾病的感染力 8](#_Toc72341036)

[2.2.2接触率矩阵 9](#_Toc72341037)

[2.2.3奇异值分解 9](#_Toc72341038)

[2.3本章小结 10](#_Toc72341039)

[第3章 基于流感监控数据的新冠肺炎疫情发展趋势预测模型 11](#_Toc72341040)

[3.1前提条件及假设 11](#_Toc72341041)

[3.2模型框架 13](#_Toc72341042)

[3.3本章小结 14](#_Toc72341043)

[第4章 实验分析及讨论 15](#_Toc72341044)

[4.1数据集 15](#_Toc72341045)

[4.2实验方案 15](#_Toc72341046)

[4.2.1数据预处理 17](#_Toc72341047)

[4.2.2计算流感和新冠的感染力 18](#_Toc72341048)

[4.2.3计算流感的接触率矩阵 19](#_Toc72341049)

[4.2.4建立感染力和新增感染人口的关系模型 20](#_Toc72341050)

[4.2.5预测新冠肺炎发展趋势 21](#_Toc72341051)

[4.3实验结果及分析 22](#_Toc72341052)

[4.3.1评价指标 22](#_Toc72341053)

[4.3.2预测结果和真实值的相关性分析 23](#_Toc72341054)

[4.3.3流感感染力和新冠肺炎感染力的相关性分析 25](#_Toc72341055)

[4.3.4流感接触率矩阵和新冠肺炎接触率矩阵相关性分析 27](#_Toc72341056)

[4.3.5更换数据 28](#_Toc72341057)

[4.4本章小结 31](#_Toc72341058)

[第5章 总结与展望 33](#_Toc72341059)

[参考文献 35](#_Toc72341060)

[致 谢 38](#_Toc72341061)

基于流感监控数据的新冠肺炎预测方法研究

刘炳辰

西南大学计算机与信息科学学院 软件学院，重庆 400715

摘 要：全球大流行COVID -19已蔓延至6大洲，并可能在全球范围内持续数年。为了帮助卫生政策制定者获得应对时间以减轻Covid-19的威胁，对疫情进行预测至关重要。基于以下观察，本文提出了利用流感监测数据预测COVID-19的方法:(1)两种疾病都是通过接触传染性呼吸道飞沫传播的，即它们可能具有相同的潜在接触结构;(2)流感比COVID-19传播速度更快，因为它的平均潜伏期和连续感染病例之间的间隔时间更短;(3)包括美国在内的许多国家已经开发了有效的流感监测系统，但针对COVID-19的监测系统仍在进展中。本文提出的预测方法首先计算流感的动态接触模式。利用所揭示的动态接触模式对同期人群感染COVID-19风险进行评估，并探索出一套基于已知传染病预测未知传染病发展趋势的方法框架。

关键词：新冠肺炎；接触率矩阵；感染力

**COVID-19 prediction method based on influenza surveillance data**

LIU Bingchen

School of Computer and Information Science, Southwest University, Chongqing 400715, PR China

Abstract: The global pandemic COVID-19 has spread to six continents and is likely to persist globally for several years. To help health policy makers get the time to respond to mitigate the threat of COVID-19, outbreak prediction is critical. A method for predicting COVID-19 using influenza surveillance data is proposed based on the following observations :(1) Both respiratory diseases are transmitted by contact with infectious respiratory droplets, i.e., they may have the same underlying contact structure; (2) Influenza spreads faster than COVID-19 because the average incubation period and the interval between successive cases of infection are shorter; (3) Many countries, including the United States, have developed effective influenza surveillance systems, but surveillance systems for COVID-19 are still in progress. The prediction method proposed in this paper first calculates the dynamic exposure pattern of influenza. The revealed dynamic exposure patterns were used to assess the risk of COVID-19 infection in the corresponding population, and a methodological framework was developed to predict the development trend of unknown infectious diseases based on known infectious diseases.

**Key words:** COVID-19; Contact matrix; Force of infection

# 第1章 导论

本章首先通过介绍新冠肺炎疫情的严重性，来阐述预测新冠肺炎疫情发展模式的必要性。简要介绍和分析目前主流的预测新冠肺炎发展趋势方法，总结现有方法的不足之处。新冠疫情爆发初期针对新冠肺炎的监测体系尚不完整，但通过分析新冠数据来预测疫情的发展趋势需要大量且真实的数据。思及于此，本文将探索能否用流感监测体系来预测新冠疫情的发展模式。在这一章中，会简单介绍了本文的主要工作；并在本章最后介绍本文的主要结构。

## 1.1研究背景

2019年12月8日第一例新冠肺炎患者在武汉确诊，2020年1月下旬新冠肺炎疫情迅速蔓延至全国各地，2020年2月11日世界卫生组织正式将其命名为“COVID-19”[1]。面对未知且来势汹汹的新冠肺炎疫情，党和政府始终把人民生命安全和身体健康放在第一位，果断采取最全面、最严格、最彻底的防控措施。经过政府和全国人民的不懈努力，国内新冠疫情已得到了全面和有效的控制[2]。然而，海外国家的COVID-19疫情传播情况并不乐观[3,4]。截至北京时间7月19日10时30分，全球累计确诊病例14310278例，累计死亡病例599 145例；美国累计确诊病例和累计死亡病例分别为3833 271例和142877例，是全球疫情最严重的国家[5]。COVID-19的爆发给公众的健康带来了严峻的挑战[6]。

预知疫情的发展趋势可提前制定应对措施，控制疫情的发展。在疾病爆发初期，没有疫苗和治疗药物，政府的关注点在于如何制定隔离策略[7]。使用数学模型估计人群covid-19感染率有助于了解疾病的流行率，便于政府及时出台应对政策[8]。挖掘疫情的内在发展规律、预判未来走势对于控制疫情具有重要的实践意义[9]。从新冠肺炎爆发之初，国内外的学者就开始进行确诊人数分析和趋势预测，并发表了大量的学术论文。

SEIR是经典的传染病模型之一[10]。国内外各个团队对疫情趋势预测展开研究，采用了SEIR模型、改进的SEIR模型，但预测结果存在较大浮动[11]。因为该模型中的传染率为常数，难以对新冠肺炎传播建立准确模型并完成传播趋势的预测。

除了传统的传染病模型，学者们对于新冠疫情的预测还采用了：机器学习方法[12,13]、时间序列模型[14]、使用社交媒体和网络新闻挖掘来预测covid-19的爆发[15]、结合症状预测可能感染的模型[16]等。

以上方法均需要大量的且真实的新冠数据[17]。然而在新冠疫情爆发初期，研究人员对于新冠病毒的了解较少，新冠病毒的传播模式未知[18]；关于新冠的监测体系在迅速建立但不能迅速达到理想效果[19]。

此外，有研究表明：老年人有更大的风险感染新冠肺炎，然而基于年龄结构的新冠肺炎发展趋势预测的相关工作很少[20]。

## 1.2主要工作

基于以下现象，本文旨在探索能否借助流感的监测数据来反映新冠肺炎的传播情况：1.流感和新冠肺炎感染症状类似[21,22,23]；2.流感和新冠肺炎均主要是由于人群的密切接触而传播的[21,22,23]；3.流感感染以后的潜伏期要远远小于新冠肺炎，并且没有证据表明新冠肺炎患者不存在长期潜伏的可能性[21,22,23]；4.流感患者中几乎没有无症状感染者，而这种现象在新冠肺炎患者中却普遍存在[21,22,23]；5.流感作为周期性疾病，各个国家已经建立了成熟的流感监控体系[24]。

本文假设流感和新冠肺炎具有类似的传播模式，根据按年龄划分的流感患者数据，推导出人群中的接触模式，然后建立由于人群间密切接触而产生的疾病传播模型，进而预测新冠肺炎的发展趋势。

本文提出了基于流感监控数据的新冠肺炎疫情发展趋势预测模型，并在此基础上对于采用一种传染病监控数据预测同期另一种传染病发展趋势的框架建立上做了初步的探索。

## 1.3本文结构

本文第1章为导论部分，简单介绍了新冠疫情的现状、COVID-19病毒的危害以及进行新冠预测的意义。简单说明了一些现有研究方法，通过分析现有方法的不足，并结合目前缺少按年龄划分的疫情预测，提出本文的创新点：用按年龄划分的流感数据挖掘新冠在不同年龄段人群中的发展模式。

第2章介绍目前预测新冠肺炎传播风险的主要方法，并介绍本文将会使用到的知识，为之后流感预测新冠肺炎模型的建立做铺垫。

第3章介绍本文提出模型需要的前提条件并做出了一些假设，详细介绍了流感预测新冠模型的建立步骤。

第4章详细介绍实验过程并进行结果的分析和讨论。通过修改数据的时间范围，证明本文提出模型能够预测新冠肺炎的发展趋势。通过分析实验结果得出了两个结论：1.感染力相关系数的大小决定了模型的泛化性能好坏。2.接触率矩阵相关系数决定了能否进行预测。这两条结论对于建立用一种传染病的监控数据预测另一种疾病的方法框架有重要意义。

第5章总结了本文提出模型的优缺点，并对未来的工作做了简单介绍。

## 1.4本章小结

COVID-19的爆发给全球带来了极大的挑战[6]，面对新冠肺炎的肆意蔓延，准确预测新冠疫情的发展规律，不仅能够为科学防控、减轻疫情对社会经济的影响提供科学决策，而且对定量的评估疫情所处的阶段、管控措施的效果、预测未来走势以及如何进行防控具有重要的实践意义[9]。目前对于新冠肺炎传播风险预测的主要方法有传染病模型、时间序列分析以及利用机器学习方法进行统计建模等。尽管不同预测方法均能够达到较好的预测效果，但其中仍然存在一些问题，并且需要大量的数据支撑[17]。考虑到目前对于按年龄划分的新冠传播风险的预测工作开展较少，综合新的疫情爆发初期时医疗机构没有完善的监测体系、统计数据不全、预测模型复杂等因素，本文另辟蹊径，探索根据成熟的流感监测体系获取的流感数据来分析和预测同时期新冠肺炎的传播趋势，并取得了一些成果。本文还提出了一种全新的传染病预测方法框架：利用已知的传染病监测数据预测同时期另一种传染病的发展趋势。

# 第2章 相关研究和技术方法概述

本章主要介绍关于新冠肺炎传播风险预测的现有方法和相关研究，以及实现本文提出的模型需要的技术知识。

## 2.1相关研究

目前解释与预测covid-19爆发情况的方法主要有：传染病模型[10]、机器学习方法[12,13]、时间序列模型[14]、使用社交媒体和网络新闻挖掘来预测covid-19的爆发[15]、结合症状预测可能感染的模型[16]。

在动力学模型中， SEIR及其拓展模型备受学者们的青睐[25]。同时，一些学者从统计建模的角度对新冠疫情的国内外发展趋势进行了研究，比如机器学习、线性模型和二次函数模型以及指数增长模型等。特别地，Carcione等[26]根据自然增长规律提出了一种简洁、实用的流行病动态传播率计算方法用来代替复杂不确定的基本传染数Ro。经典的SEIR模型泛化适用于各类疫情，但须考虑人员流动的感染。文献[27]采用修正的SEIR模型对COVID—19进行拟合分析预测，但由于未考虑防控措施对人员流动的影响，预测结果与国家卫健委报告人数存在较大偏差。无论是SEIR模型，还是修正的SEIR模型，均需要分析大量的参数，包括R0、恢复率（移出率）等，而人员网络存在大量的不确定性因素。

SEIR模型中，S指易感状态，E指潜伏状态，I指感染状态，R指康复或者死亡。假设一个易感状态在单位时间里与感染个体接触并被传染的概率为，整体以单位时间概率转化为感染个体，感染个体数目由潜伏群体转化而来，同时以单位时间概率转化为移除状态。

## 2.2相关技术

## 2.2.1计算疾病的感染力

感染力是用来衡量人群感染某种疾病的风险，感染力越大，人群感染这种疾病的风险就越高。Edmunds[27]等人提出了根据催化模型计算感染力的方法：

表示年龄段i易感人群感染的风险，又称为感染力[15]， 表示年龄段i的年龄跨度，Ci是年龄类别i的病例报告数。

Rohani[28]等人提出一种基于接触率矩阵的感染力计算方法：

表示年龄段i易感人群和年龄段j感染人群接触后被感染的概率。是接触矩阵，表示一个i年龄段的人会和多少个j年龄段的人接触。表示年龄段j的感染人口，表示年龄段j的样本人口。

## 2.2.2接触率矩阵

接触率矩阵用来表示人群的接触情况，记录了每个年龄段中的每个人与其他各个年龄段人群接触的均值。如是接触矩阵第i行j列的值，表示一个i年龄段的人会和多少个j年龄段的人接触。

将公式（2-2）、（2-3）联立，可以计算接触率矩阵。联立会产生一个欠定方程组。具体来说，将公式（2-3）调整为矩阵方程Ax=B的形式（见公式2-4），接触矩阵用规格为1\*16的列向量x表示，作为规格为4\*16的参数矩阵A，感染力为常数项B。

## 2.2.3奇异值分解

假设M是一个m×n阶矩阵，其中的元素全部属于域 K，存在一个分解使得

公式（2-5）中的U是m×m阶酉矩阵；Σ是半正定m×n阶对角矩阵；V\*是V的共轭转置。常见的做法是为了奇异值由大而小排列。如此Σ便能由M唯一确定了。在矩阵M的奇异值分解中,U的列组成一套对M的正交"输入"或"分析"的基向量。这些向量是MM\*的特征向量。V的列组成一套对M的正交"输出"的基向量。这些向量是M\*M的特征向量。Σ对角线上的元素是奇异值，可视为是在输入与输出间进行的标量的"膨胀控制"。这些是M\*M及MM\*的奇异值，并与U和V的列向量相对应。

## 2.3本章小结

目前预测新冠肺炎传播风险的几种方法是传染病模型、时间序列分析以及机器学习。本章介绍了一些本文模型需要的预备知识。基于这些知识逐步建立起了用流感监测数据预测新冠肺炎传播风险的模型。

# 第3章 基于流感监控数据的新冠肺炎疫情发展趋势预测模型

本章将具体介绍本文提出的基于流感监控数据的新冠肺炎传播风险预测模型。首先介绍建立模型的前提条件及模型中的假设，然后通过一个框架图直观地介绍模型的预测方法，最后具体说明模型是如何根据上一章的相关技术建立起来并进行预测的。

## 3.1前提条件及假设

据统计，每年有数百万美国人患上流感[29]， 流感是由流感病毒引起一种传染性呼吸道疾病。引起流感的病毒和新型冠状病毒是不同的，但流感和新冠肺炎症状重叠；COVID-19的最常见症状是：发热、干咳、疲劳；流感的一些症状包括发烧或感觉发热/发冷、咳嗽、疲乏（疲劳）等[21,22,23]。流感疾病和新冠病毒均通过接触、飞沫和污染物传播[23]。因此相同的公共卫生措施，如保持手部卫生和良好的呼吸礼仪，可以缓解二者的传播[22]。考虑到流感的中位潜伏期和连续病例之间的时间间隔比新型冠状病毒肺炎更短[23]，流感的监测体系比新冠更成熟，在本文中，假设流感和新冠肺炎在人传人中具有相同的传播方式，分析能否通过同期的流感数据来预测新冠肺炎的流行状况。有关研究表明，不同年龄段人群感染新冠肺炎的风险不同[10]，因此我们根据按年龄划分的流感数据，计算出人群的接触矩阵，再将接触矩阵应用到新冠肺炎感染率的预测上。

Edmunds[27]等人提出了一种计算感染力的方法，本文采用这种方法计算流感对于不同年龄阶段人群的感染力，再根据Rohani[28]等人提出接触率矩阵和感染力的关系公式计算人群的接触率矩阵。根据以下研究，本文流感预测新冠的模型将只考虑接触率矩阵、接触后感染概率、人群情况三个因素，而不考虑已感染人群的恢复情况：

1.预测由SARS-CoV-2引起的COVID-19大流行过程的关键问题之一是免疫反应保护宿主免受再感染的程度和持续时间。Edridge等人[30]的研究表明：对于季节性冠状病毒，保护性免疫是短暂的。2.Roda等人[17]通过使用Akaike信息准则（AIC）进行模型选择，表明SIR模型在表示已确认病例数据中包含的信息方面比SEIR模型表现得更好。这表明复杂模型在进行疾病预测时并不会比简单模型更可靠。因此我们将模型简单化，忽略了免疫带来的影响。

考虑到新冠肺炎有较长的潜伏期，新增的感染人口未必会在下一周表现出症状，因此本文在了解了新冠肺炎的潜伏期分布情况，假设潜伏期服从均值为5，标准差为3.3的正态分布，并假设所有的感染者都会在三周内表现出感染症状，即能够被检测出来。利用累积分布函数将新增感染率按照假设的分布分配到未来的几周。

获取到的流感、新冠数据均为抽样数据，本文假设这些抽样数据可以代替总体，如样本数据的感染率可以代替美国总体人口的感染率。

## 3.2模型框架

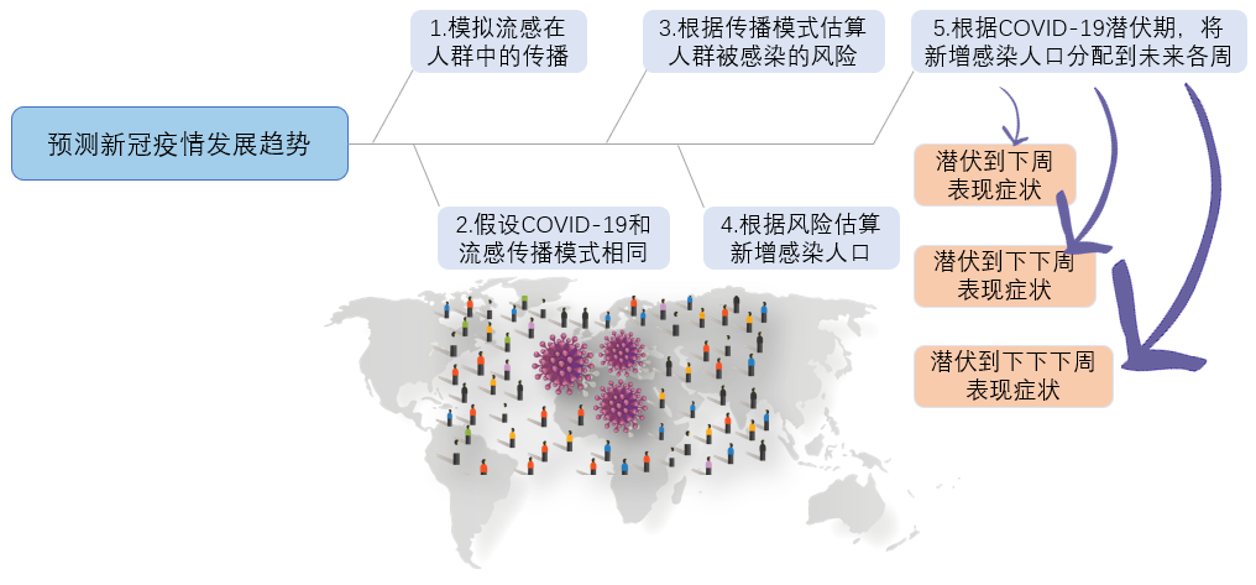


图3.1预测流程图

Fig.3.1 Forecast flow chart

1.根据按年龄划分的流感数据计算人群每周的接触率矩。

2.将流感数据计算的接触率矩阵近似看作导致新冠病毒传播的人群接触率矩阵，即用流感数据计算的接触率矩阵模拟同时期新冠病毒在人群中的传播情况。

3.根据模拟的新冠病毒的传播情况估算对应时间段各个年龄阶段人群感染新冠肺炎的风险，即用1中计算的接触率矩阵估算同时期人群感染新冠肺炎的风险。

4.通过统计分析，建立感染风险与新增感染人数之间的转化关系。进而将估计的人群感染新冠肺炎的风险转化为新增感染人口。

5.考虑到新冠肺炎具有较长潜伏期，根据新冠肺炎潜伏期的统计分布模型，将4中计算得到的新增感染人口按概率分配到未来的各周。比如，如果计算得到了由于第10周人群密切接触而产生的新增感染人口，这个新增感染人口可能会在未来的11周、12周……表现出症状并被统计到，因此需要将新增感染人口按照新冠肺炎潜伏时间的累积分布函数分配到未来的几周中。

6.不考虑患者恢复情况，将新增感染人口加到上一周的感染人口上，完成新冠肺炎发展趋势的预测。

## 3.3本章小结

本章介绍了本文流感预测新冠模型的前提条件、假设以及建立步骤。流感和新冠在传播模式、感染症状等方面高度一致；本文提出的模型不考虑患者的恢复情况，并假设新冠肺炎的潜伏期服从正态分布。预测模型建立方法可以概括为：首先根据流感数据计算出流感的感染力，然后计算出人群的接触率矩阵。假设新冠肺炎的传播情况和流感相同，根据人们的接触情况以及感染者在人群中的概率，计算出本周由于人群密切接触而可能产生的新增感染人口。再根据流新冠的潜伏期，预测未来几周的新冠肺炎发展趋势。具体实验过程将在第四章介绍。

# 第4章 实验分析及讨论

本章将具体介绍实验的过程，数据集的选取，模型的建立以及最后的仿真。此外，还给出了给出判断一种疾病数据能否预测另一种疾病数据的方法。为了评估模型的泛化性能，本章使用时间序列分析进行了新冠风险的预测，将结果和流感预测新冠模型的预测结果对比。最后进行了关键参数的讨论。

## 4.1数据集

从美国疾控中心网站获取到了美国公共卫生实验室和临床实验室2019、2020年流感和新冠按年龄划分的数据集。该数据集统计了0-4岁、5-24岁、25-64岁以及65岁以上人群感染各种流感的样本数据；和0-4岁、5-17岁、18-49岁、50-64岁以及65岁以上人群感染新冠肺炎的样本数据。

从世界银行公开数据库的健康营养与人口统计数据集中获取到美国2020年的人口数据。该数据集统计了0岁到64岁每一个年龄段的美国人口数据以及65岁以上的美国人口数据。

## 4.2实验方案

个体接触模式在时间上存在差异，本文通过分析不同时期的流感患者数据，计算出流感在人群中的动态传播模式和动态接触矩阵，利用动态接触矩阵预测新冠肺炎的新增感染率。

数据的年龄结构：流感和新冠不同，需对流感数据进行调整。在抽样数据是在美国全体人口均匀采样的前提下，假设感染人口的年龄分布和美国总人口的年龄分布相同，将流感感染数据的年龄结构调整为与新冠数据一致。

数据的时间间隔：美国公共卫生实验室统计了每周不同年龄段人口累计感染流感的情况，考虑到新冠肺炎的潜伏期参差错略，本文假设每周统计到的数据包含本周接触感染者并在本周内表现症状的人口。

数据的起止时间选取：考虑到流感的季节性特征[21]以及新冠疫情的阶段性发展特征[22]，以及新冠数据的起始日期是2020年第10周，本文选取2020年第10-20周数据进行分析。将2020年第10周及以后的时间点孤立出来作为疾病传播预测的起点，假设所有的新增感染者都是通过第10周人群的接触传播感染的。

接触模式预测：本文旨在计算一种传染病在人群中的传播模式，预测另一种传染病的发展情况，提出一种较为通用的预测方法，需要考虑传播模式的相关性，即无论两种疾病传播模式是否类似，都能对目标疾病进行准确预测。因此，本文采取的方案是用不同周期长度的解释变量疾病的接触矩阵的期望作为响应变量疾病的接触矩阵。

新冠肺炎潜伏期分布：目前普遍认为的covid-19的潜伏期为1-14天，平均潜伏期为5天，97.5%的感染者会在11.5天内出现临床症状[1]。基于以上发现，本文假设新冠肺炎的潜伏期服从期望为5，标准差为3.3的正态分布。

新增感染人数与感染力的关系讨论：相关研究表明，流感的潜伏期多为2~4天，流感患者一般会在一周内恢复[21,22]，本文假设1.如果易感染群在本周感染了流感，那么他会在本周表现出症状；2某一周的流感感染人口全部是由于上一周人群的密切接触导致的。基于上述假设，可以通过探讨第i周流感感染力和第i+1周流感感染人数之间的关系近似得到感染力和新增感染人口的关系。在计算新冠肺炎的新增感染人口时采用根据流感数据计算的感染力和新增感染人数的关系式。

## 4.2.1数据预处理

假设感染数据中病人的年龄分布和美国总体人口的年龄分布相同，根据美国人口的年龄分布调整流感数据的年龄分布。为了减少模型的误差，仅对流感和新冠数据中年龄结构不同的数据进行调整。

获取到的流感数据有0值，表示对应周并未有人感染。考虑到对covid-19的防控措施也会影响到流感，我们假设数据反映的是真实的情况，对0值数据未进行填充操作。

处理结果如图4.1所示，其中H1、U2S（Unable to Subtype）、H3、H1N1（H1N1pdm09）、SnP（Subtyping not Performed）、VL（Victoria Lineage）、YL（Yamagata Lineage）、LU（Lineage Unspecified）、H3N2（H3N2v）是各种流感子型，在本文剩余部分依然按照以上简称形式称呼各种流感子型；All Flu表示所有子型流感的数据。

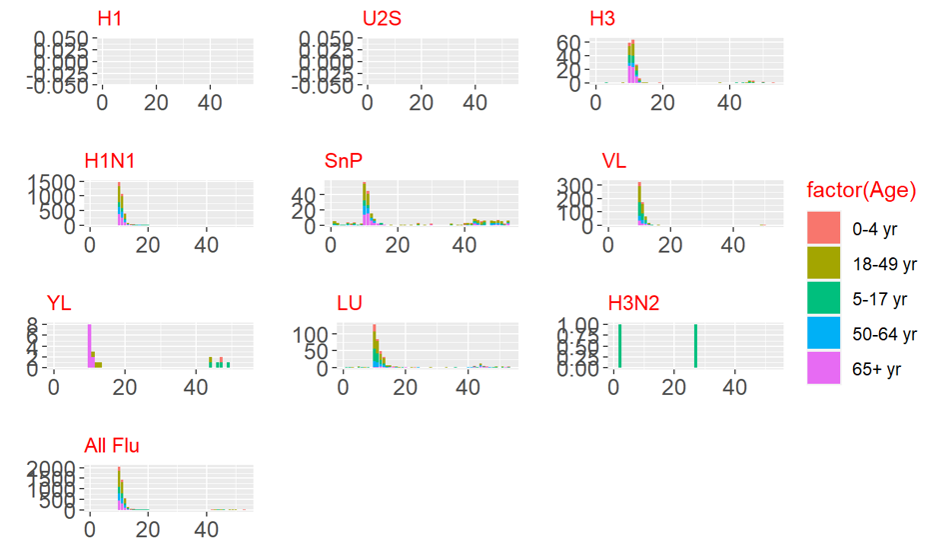


图4.1 2020年按年龄划分的流感患者数据

Fig.4.1 Influenza data by age in 2020

数据是十分稀疏的，并且每种流感并不都所有年龄分组的患者，如第10周YL数据就只有65+以上人群被感染。这种现象会给感染力的计算带来比较大的困难。

无论是新冠肺炎数据还是临床实验室的流感数据，每周进行测试的样本人数都不是固定的。本文假设样本数据表现的统计学特征可以代表美国总体人口，根据样本数据的感染者数量占样本人口的比例，将样本感染人口调整为总体的感染人口。

## 4.2.2计算流感和新冠的感染力

根据公式（2-2）计算2020年第10-20周各种流感和新冠肺炎的感染力，由于公式的局限性，无法计算最后一个年龄组的感染力，在本文剩余部分将不会对65+年龄组进行分析。

正如4.2.1中所述，稀疏的数据会导致部分感染力的计算结果为Inf（表示无穷大）或者NaN（表示无法计算），如图4.2所示，其中每一列表示一个年龄分组，每一行表示一周。

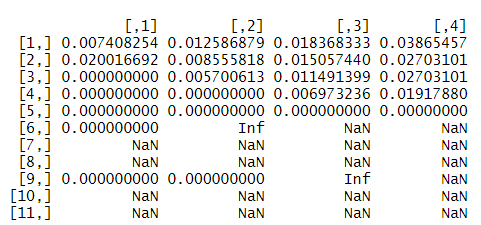


图4.2 2020年第10-20周四个年龄段SnP的感染力

Fig.4.2 Force of infection of SNP in age group 10-20 in 2020

我们将数值为Inf或者NaN的感染力替换为一个极小的数字0.0001，用来表示人群极不可能感染这种疾病。

## 4.2.3计算流感的接触率矩阵

根据公式（2-3）将4.2.2中计算的感染力代入，采用奇异值分解的方法计算2020年第10-20周各种流感的接触率矩阵以及新冠肺炎的接触率矩阵。美国疾控中心统计了公共卫生实验室的按年龄划分的流感数据，但未统计每次的样本人数，即公式（2-3）中的未知；临床实验室统计的数据有样本人数但未按年龄划分。本文将二者结合，将公共卫生实验室的按年龄划分的数据和临床实验室的样本人口合并为一个新的数据集。

相关研究表明不同年龄阶段感染新冠肺炎的概率是不同的[20]，即公式（2-3）中的q因年龄而异。本文为方便起见，假设所有年龄段易感人群接触到感染者之后被感染的概率均为0.05。

## 4.2.4建立感染力和新增感染人口的关系模型

本文假设在某一周的流感患者在下一周都会恢复，即下一周的流感感染人口都是由于上一周人群的密切接触而产生的。认为下一周的感染人口都为新增感染人口，建立第i周感染力和第i+1周感染人数的关系模型，近似认为该模型为感染力与新增感染人口的关系。

考虑到2020年流感数据有限，不能得到一个比较准确的关系式，本文采用2019年数据探索感染力与新增感染人口的关系。2019年流感数据如图4.3所示：

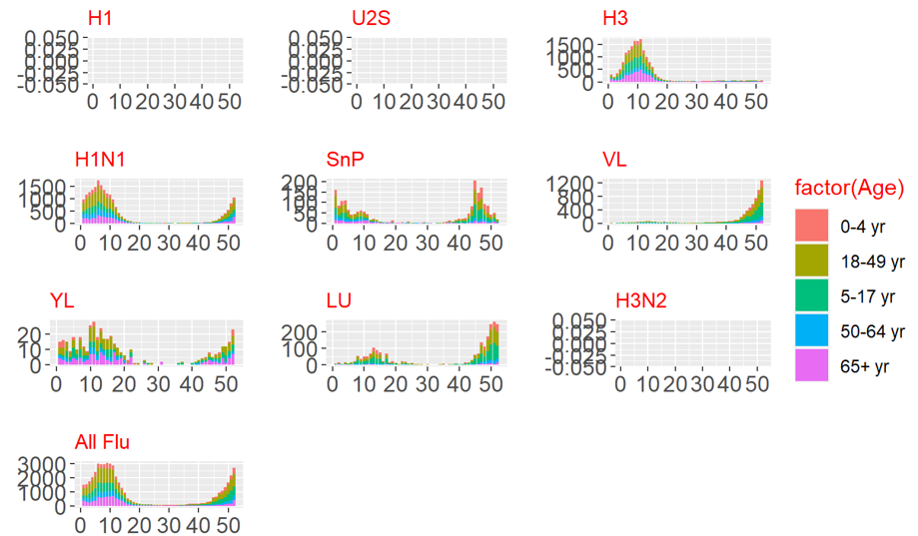


图4.3 2019年按年龄划分的流感患者数据

Fig.4.3 Influenza data by age in 2019

最后感染力与新增感染人口满足如下关系式，下标i表示第i周：

## 4.2.5预测新冠肺炎发展趋势

如果具体到一周内的每一天，想要预测新增感染者在下周表现出症状即被确认并被统计为感染者，这将会非常复杂：如果covid-19病毒在周一入侵了一位易感染者体内，想要患者下一周呈现症状，那病毒在该名易感人体内的潜伏其为7-13天，具体到每一天会有很多种情况，为了将问题简单化，本文做出如下假设和限制条件：

1.covid-19病毒在第一周接触感染者并将在第二周表现出症状的易感人群体内的潜伏期7天。

2.covid-19病毒在第一周接触感染者并将在第三周表现出症状的易感人群体内的潜伏期8天， 14天。

3.非上述两种情况的所有易感人群将在第四周呈现感染症状。

i年龄段的新增感染人口的预测公式为（4-5）：

为i年龄段易感染群接触到感染者后被感染的概率，本文假设全为0.05；

是根据流感数据计算得到的接触率；为j年龄段的感染新冠肺炎的人口；为j年龄段的总人口；表示新冠肺炎潜伏期为x天的概率，可以通过正态分布的累积分布函数得到；*Q*是将感染力转换为新增感染人口的系数，通过4.2.4得到。该公式*Q*之前的部分实际上是在计算新冠肺炎对于各个年龄段人群的感染力，只不过是根据流感数据计算得到的接触率而并非是通过新冠数据计算得到的。

预测第11周的新增感染人口，只需要第10周的数据，并且只考虑潜伏期小于7天的情况。预测第12周的新增感染人口，需要考虑到由于第10周人群密切接触而产生的新增感染人口在两周之后表现出症状的情况以及由于第11周人群密切接触而产生的新增感染人口在一周之后表现出症状的情况。预测13周及以后的新增感染人口，需要前三周共同贡献的情况。换言之，第11周的新增感染人口由第10周人群密切接触产生；第12周的新增感染人口由第11周和第10周人群密切接触产生；第13周及以后的新增感染人口由前三周人群的密切接触共同产生。

将新增感染人口叠加到前一周的感染人口完成新冠肺炎发展模式的预测。

## 4.3实验结果及分析

## 4.3.1评价指标

本文旨在预测新冠肺炎的发展趋势，而并非精确预测未来每一周的感染情况。因此评价本文提出模型性能好坏的唯一标准就是相关性分析。分析相关性，用皮尔逊相关系数衡量。在自然科学领域中，该系数广泛用于度量两个变量之间的线性相关程度。

## 4.3.2预测结果和真实值的相关性分析

有9种流感子型，本文在9种流感的基础上增加了一组将所有流感累加到一起的数据。但是H1、U2S以及H3N2的统计数据显示2020年第10到第20周并未有人感染这三种流感。因此本文可以利用7组数据进行新冠预测。根据本文提出预测模型预测的结果和真实值之间的对比如图4.4、4.5所示：

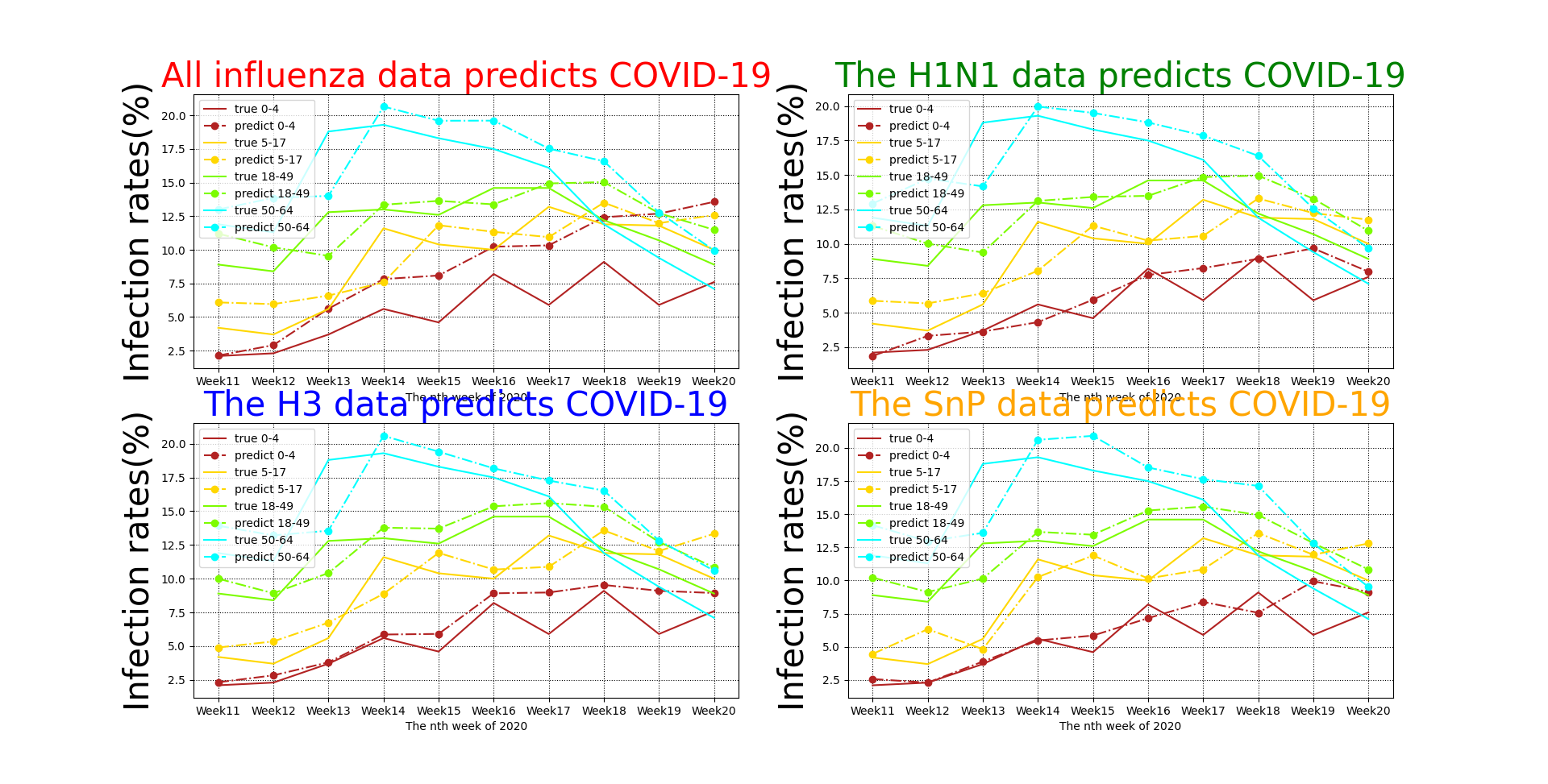


图4.4 2020年第10周到第20周预测结果（1）

Fig.4.4 Forecast results from the 10th to the 20th week of 2020(1)

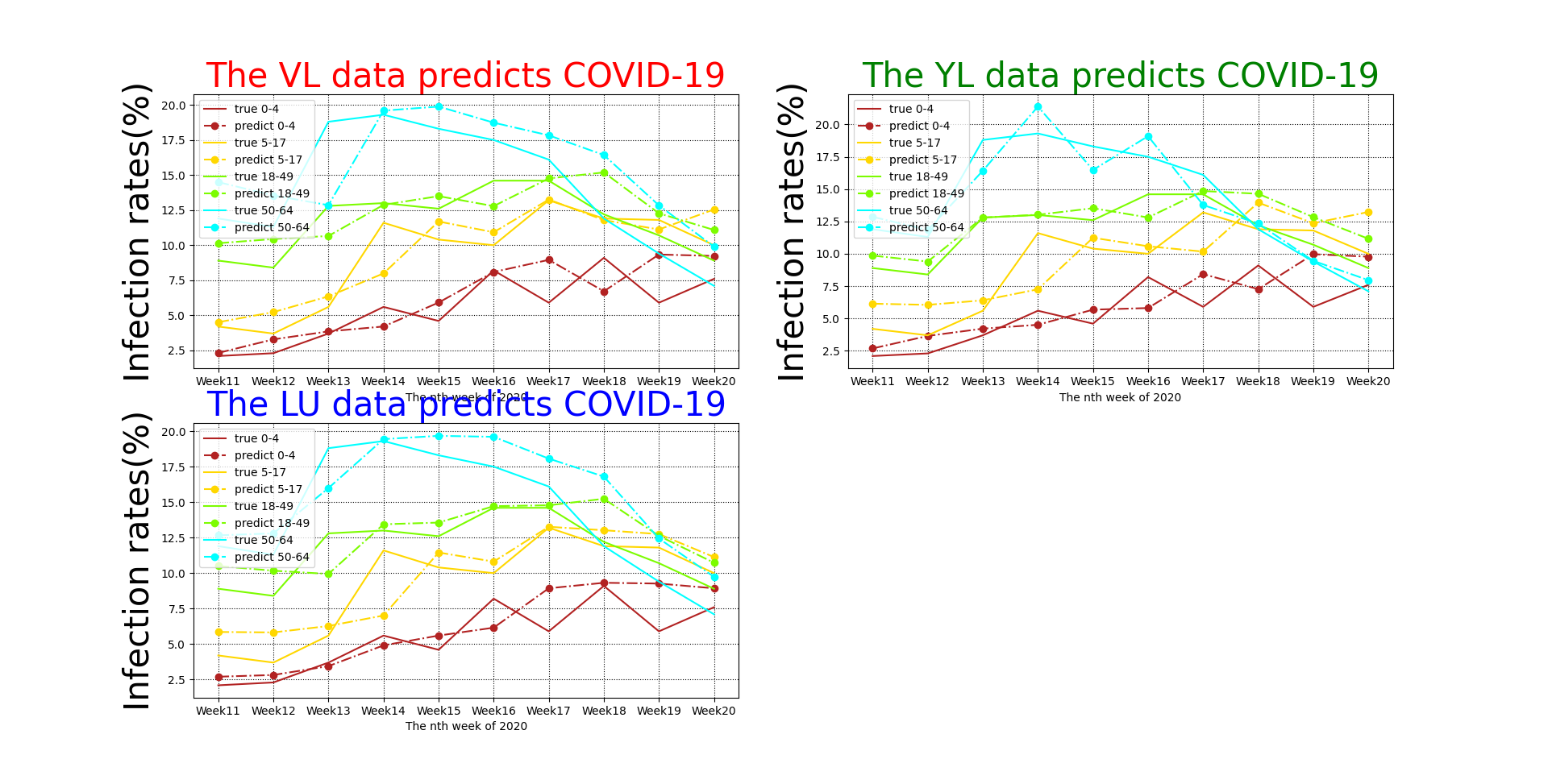


图4.5 2020年第10周到第20周预测结果（2）

Fig.4.5 Forecast results from the 10th to the 20th week of 2020(2)

横坐标表示的是第几周，纵坐标表示的是感染率。实线是各个年龄段感染率的统计值，虚线是根据本文预测模型预测的感染率。相同颜色表示的是同一个年龄阶段。通过分析结果图，我们可以发现预测的结果具有明显的滞后性，即我们的预测结果大约是前一周真实值后移一周得到的。为了得到更直观的信息，将根据各种数据预测的结果和真实值之间的相关系数可视化，结果如图4.6所示。

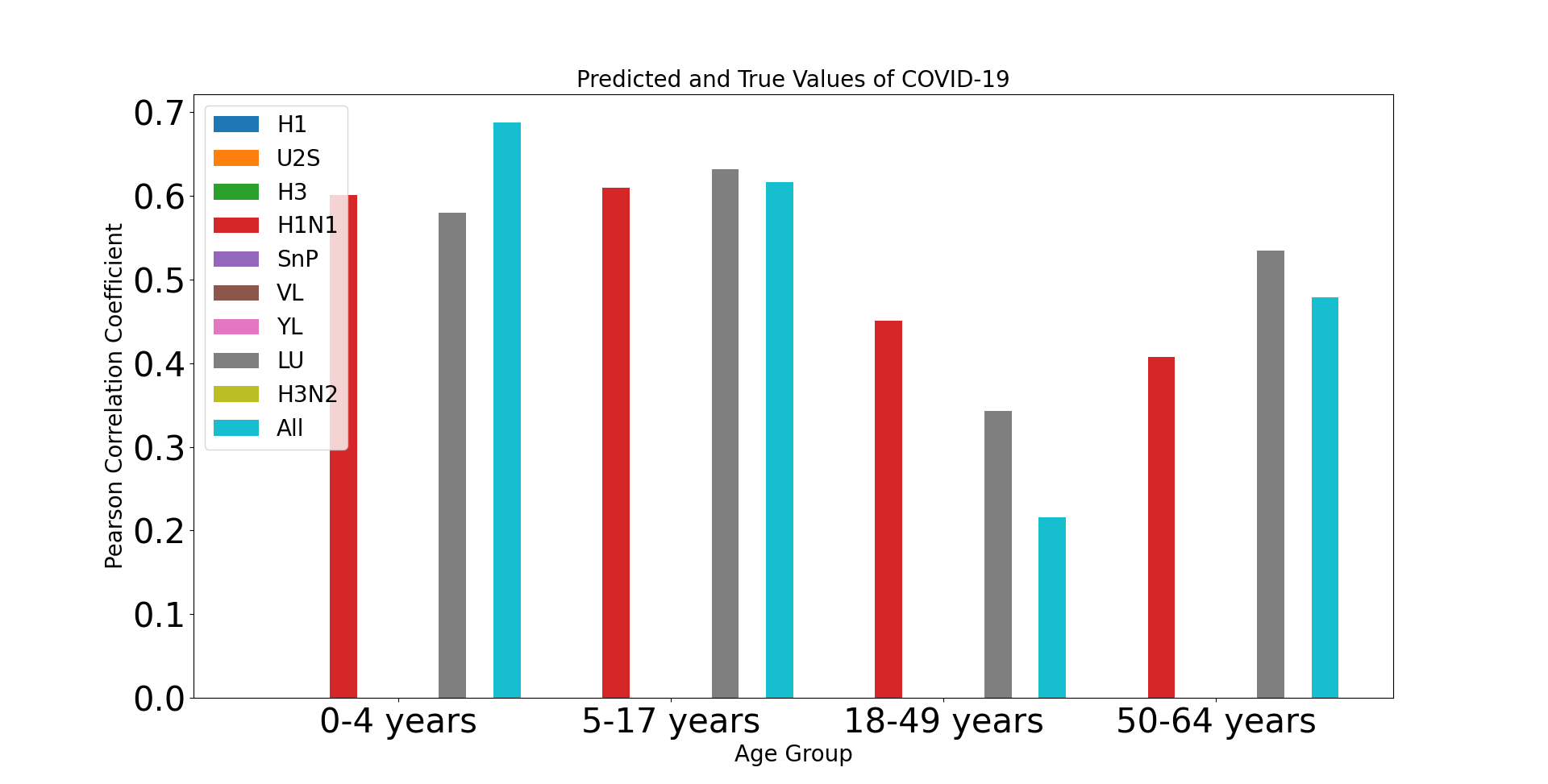


图4.6 2020年第10周到第20周预测结果和真实值的皮尔逊相关系数

Fig.4.6 Pearson correlation coefficient between predicted results and true values during the 10th to 20th week of 2020

图中显示的信息为：在四个年龄阶段均只有根据H1N1、LU、All的数据预测的结果和真实值之间是线性相关的，且相关系数均不大于0.7。0-4岁年龄组中LU、H1N1、All的预测结果和真实值之间的相关性依次递增；5-17岁年龄组中相关性按照H1N1、All、LU的顺序递增；18-49岁年龄组中H1N1、LU、All依次递减；50-64岁年龄组中H1N1数据预测结果和真实值之间的相关性小于All数据预测结果和真实值得相关性，二者均小于根据LU数据预测结果和真实值之间的相关性。

通常认为皮尔逊相关系数大于0.8表示两组数据具有强线性关系。因此我们的预测效果并不理想。尽管如此，我们依然可以通过分析流感感染力和新冠肺炎感染力的相关性、流感接触率矩阵和新冠肺炎的接触率矩阵的相关性来得到一些对于用一种疾病的监测体系来预测另一种疾病发展模式框架的有用信息。具体结果将在4.3.2和4.3.3中阐述。

## 4.3.3流感感染力和新冠肺炎感染力的相关性分析

如前所述，我们有7组数据来预测新冠肺炎的发展趋势，预测的结果和真实值之间的相关性各有差异，因此如何判断哪种疾病的预测结果和真实值之间有更高的相关性是值得我们探讨的。

在本节中将通过分析流感感染力和新冠肺炎感染力的相关性，来探讨感染力是否是决定预测效果的一个因素。各种流感和新冠肺炎感染力的相关性结果如图4.7所示。

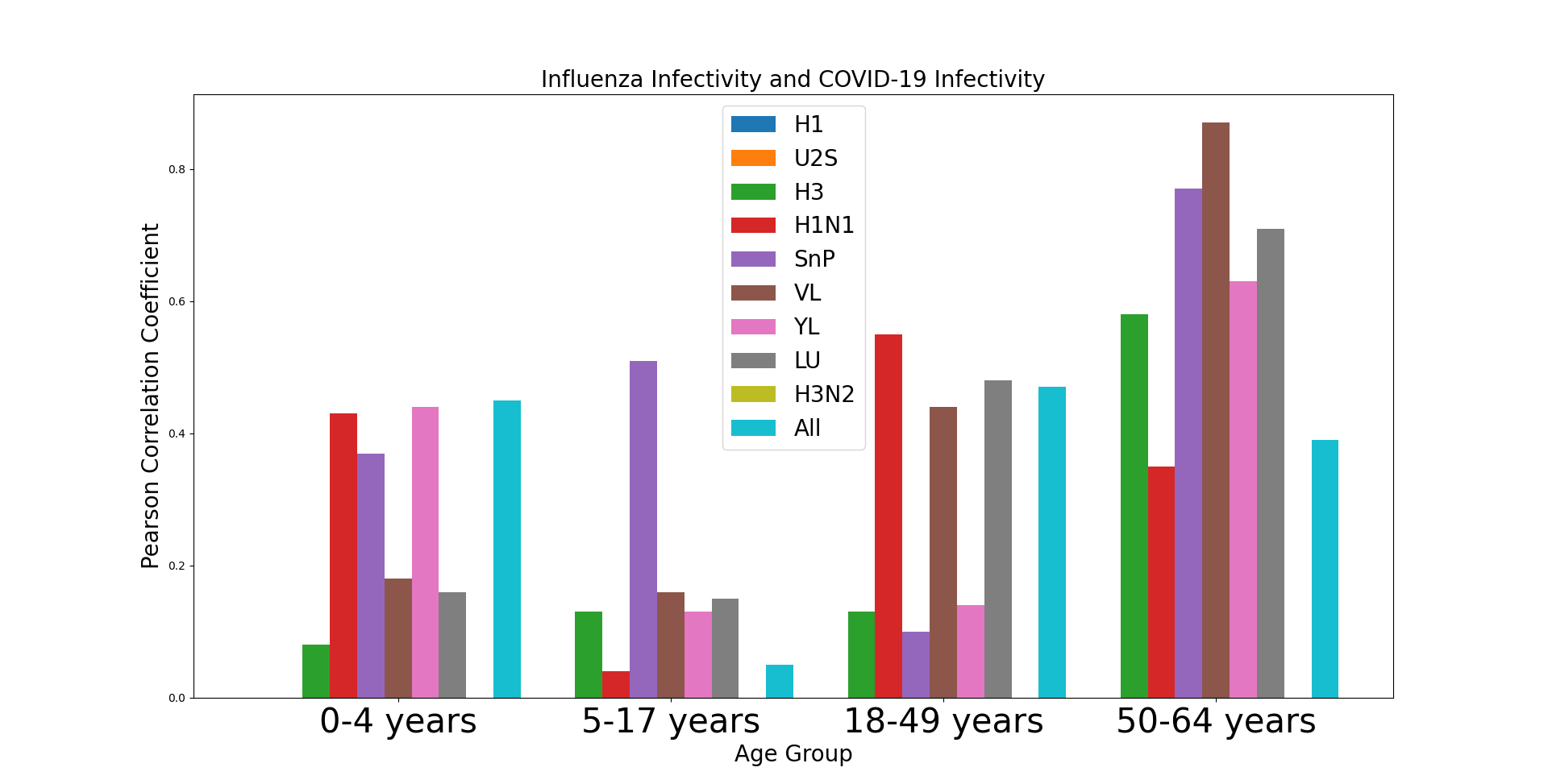


图4.7 2020年第10周到第20周流感和新冠感染力皮尔逊相关系数

Fig.4.7 Pearson correlation coefficient of influenza and COVID-19 infection from the 10th week to the 20th week in 2020

0-4岁年龄组中各种流感的感染力和新冠肺炎的感染力的相关性从小到大的顺序依次是：H3LUVLSnPH1N1YLAll；5-17岁年龄组中SnP的感染力和新冠肺炎的感染力的相关性大于VL的感染力和新冠肺炎的感染力的相关性，大于LU的感染力和新冠肺炎的感染力的相关性，大于YL的感染力和新冠肺炎的感染力的相关性，大于H3的感染力和新冠肺炎的感染力的相关性，大于All的感染力和新冠肺炎的感染力的相关性，大于H1N1的感染力和新冠肺炎的感染力的相关性；在18-49岁年龄组中相关性按照SnP、H3、YL、VL、All、LU、H1N1的顺序依次递增；50-64岁年龄组中相关性从小到大依次为：H1N1AllH3YLLUSnPVL。

0-4岁年龄组中预测值和真实值相关性从小到大依次为LU、H1N1、All，而0-4岁年龄组中感染力相关性也是按照LU、H1N1、All的顺序依次递增；5-17岁年龄组中预测结果和真实值相关性H1N1AllLU，感染力相关性H1N1AllLU；18-49岁年龄组二者也都是按照H1N1、LU、All递减；50-64岁年龄组预测和真实：H1N1AllLU，感染力H1N1AllLU。

通过以上分析，本文认为感染力的相关性可能会决定预测效果，感染力和新冠感染力越相关，预测效果可能越好。此发现将在4.3.4小节中得到验证。在这里尝试用理论解释：如果人群感染流感和COVID-19的风险越相关，二者在人群中的传播模式就愈接近，进而通过流感数据预测COVID-19发展的效果就越好，即相关性越强。

## 4.3.4流感接触率矩阵和新冠肺炎接触率矩阵相关性分析

本节和4.3.2类似，分析接触率矩阵之间的相关性对预测结果的影响。各种流感数据计算的接触率矩阵和根据COVID-19数据计算的接触率矩阵的相关性如图4.8所示。

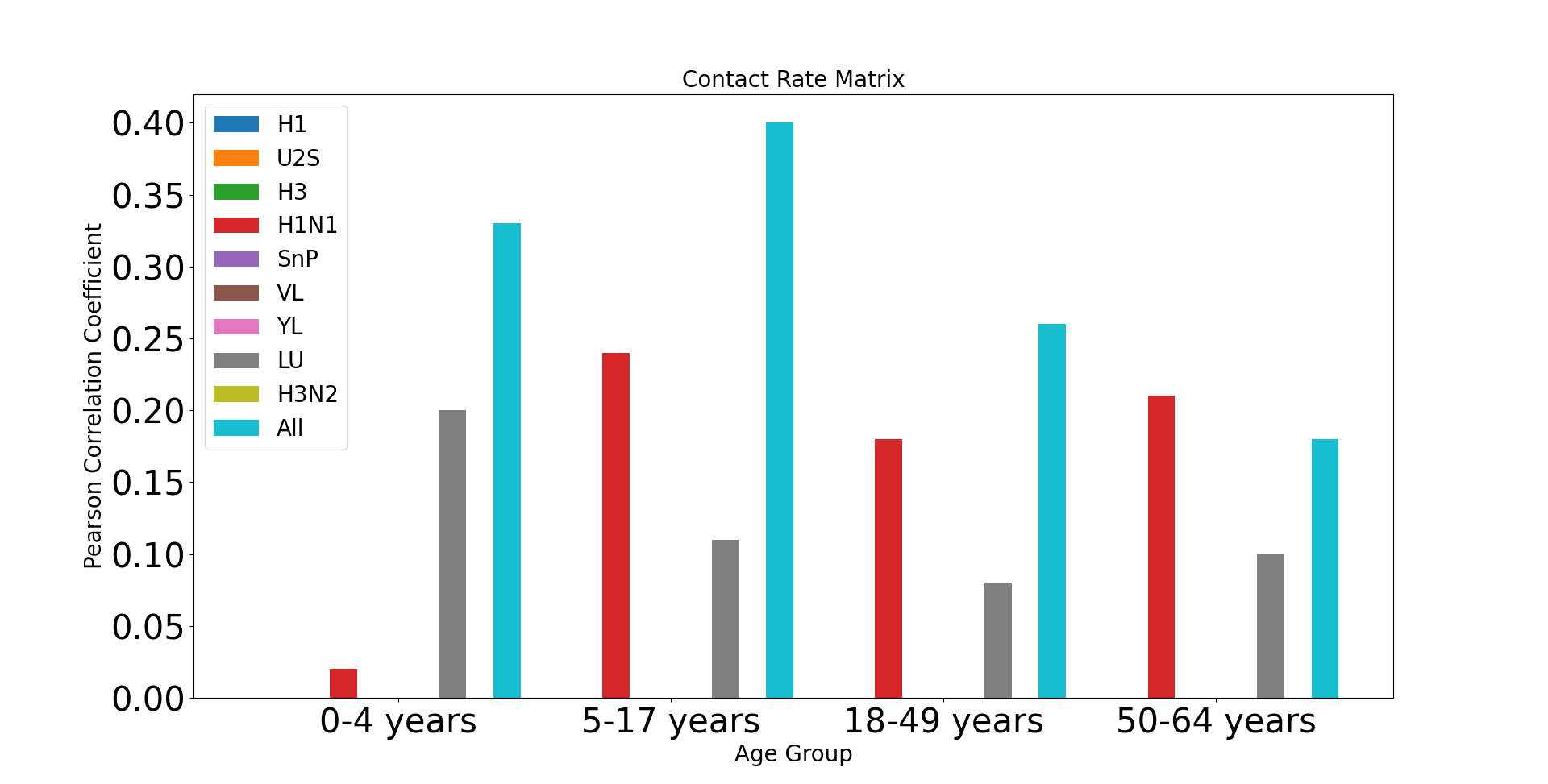


图4.8 2020年第10周到第20周流感和新冠接触矩阵的皮尔逊相关系数

Fig.4.8 Pearson correlation coefficient between influenza and COVID-19 contact matrix during weeks 10 to 20, 2020

0-4岁：H1N1LUAll；5-17岁：LUH1N1All；18-49岁：LUH1N1All；50-64岁LUAllH1N1。接触率矩阵的相关性和预测结果的相关性没有直接关系，但是接触率矩阵中有相关性的数据与预测结果和真实值之间有相关性的流感类型相同，或可以通过接触率矩阵提前筛选出可以用哪些流感数据进行预测。此发现也将在4.3.4小节中得到验证，并生成结论。

## 4.3.5更换数据

Silverman等人采用了和本文相同的COVID-19数据进行新冠疫情的预测，并证明所采用的数据不能代表人群感染新冠肺炎的真实情况，人群的感染率被严重低估了。因此，我们有理由相信COVID-19数据的不真实性影响了我们模型的预测效果。

考虑到流感具有季节周期性，并且流感季节大部分是在秋冬季节[21]。本文选择了2020年第40到48周的流感和COVID-19数据再次进行预测。结合4.3.2与4.3.3的发现：感染力相关系数决定了预测结果和真实值相关系数的大小；接触率矩阵决定了预测结果是否和真实值相关。因此我们先通过分析2020年第40周到48周流感数据计算的感染力、接触率矩阵与对应COVID-19的感染力、接触率矩阵的相关性来找到最优预测数据。

接触率矩阵相关系数如图4.9所示，能够用来预测的数据只有H1N1、SnP以及All。

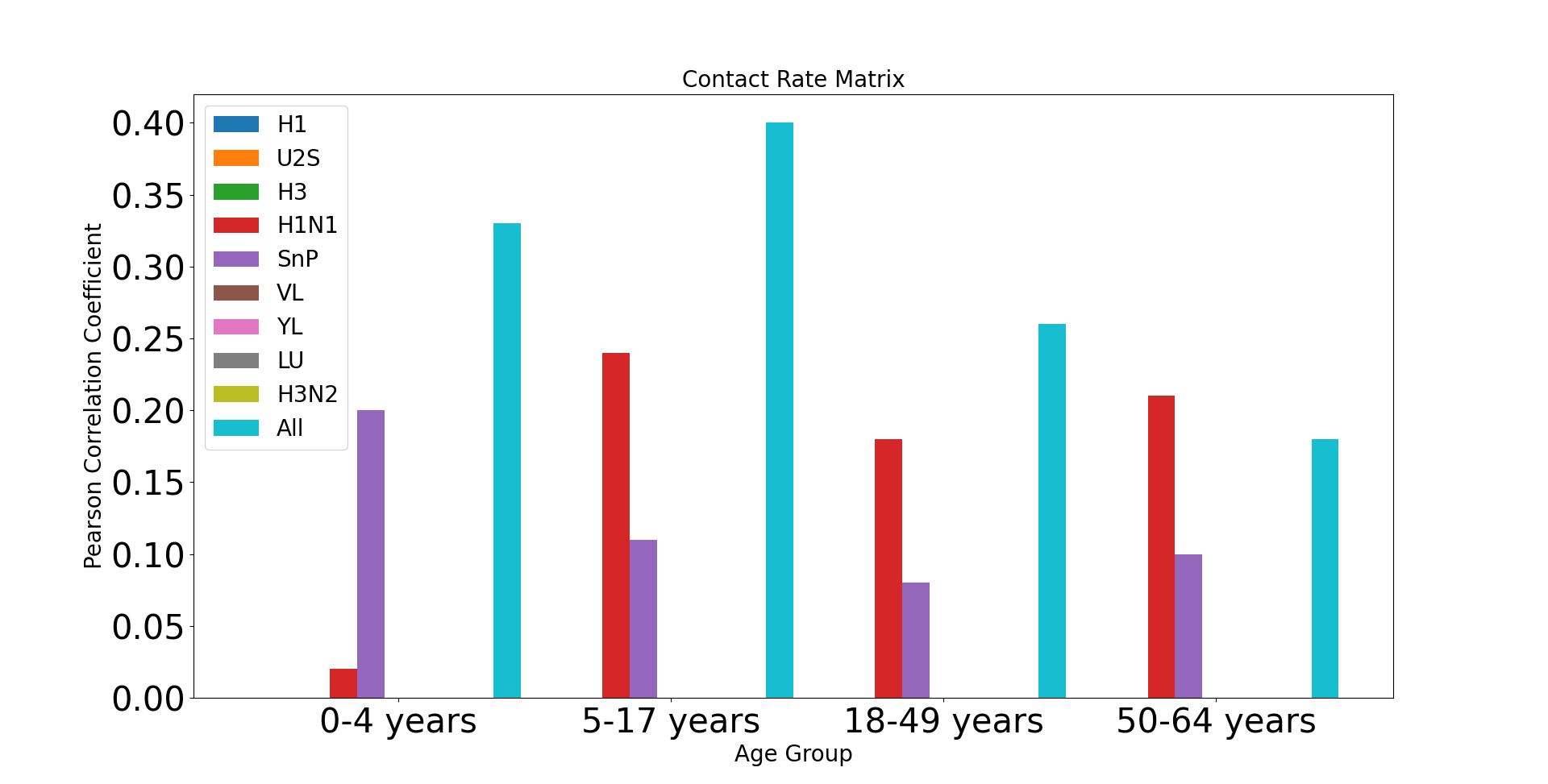


图4.9 2020年第40周到第48周流感和新冠接触率矩阵的皮尔逊相关系数

Fig.4.9 Pearson correlation coefficient of influenza and coronavirus exposure matrix from week 40 to week 48, 2020

感染力相关系数如图4.10所示， 0-4岁、5-17岁、18-49岁相关性最高的都是All；50-64岁H1N1、SnP和All中依然是all最高。因此，我们选择用All数据来预测2020年第40周-第48周的COVID-19疫情发展趋势。

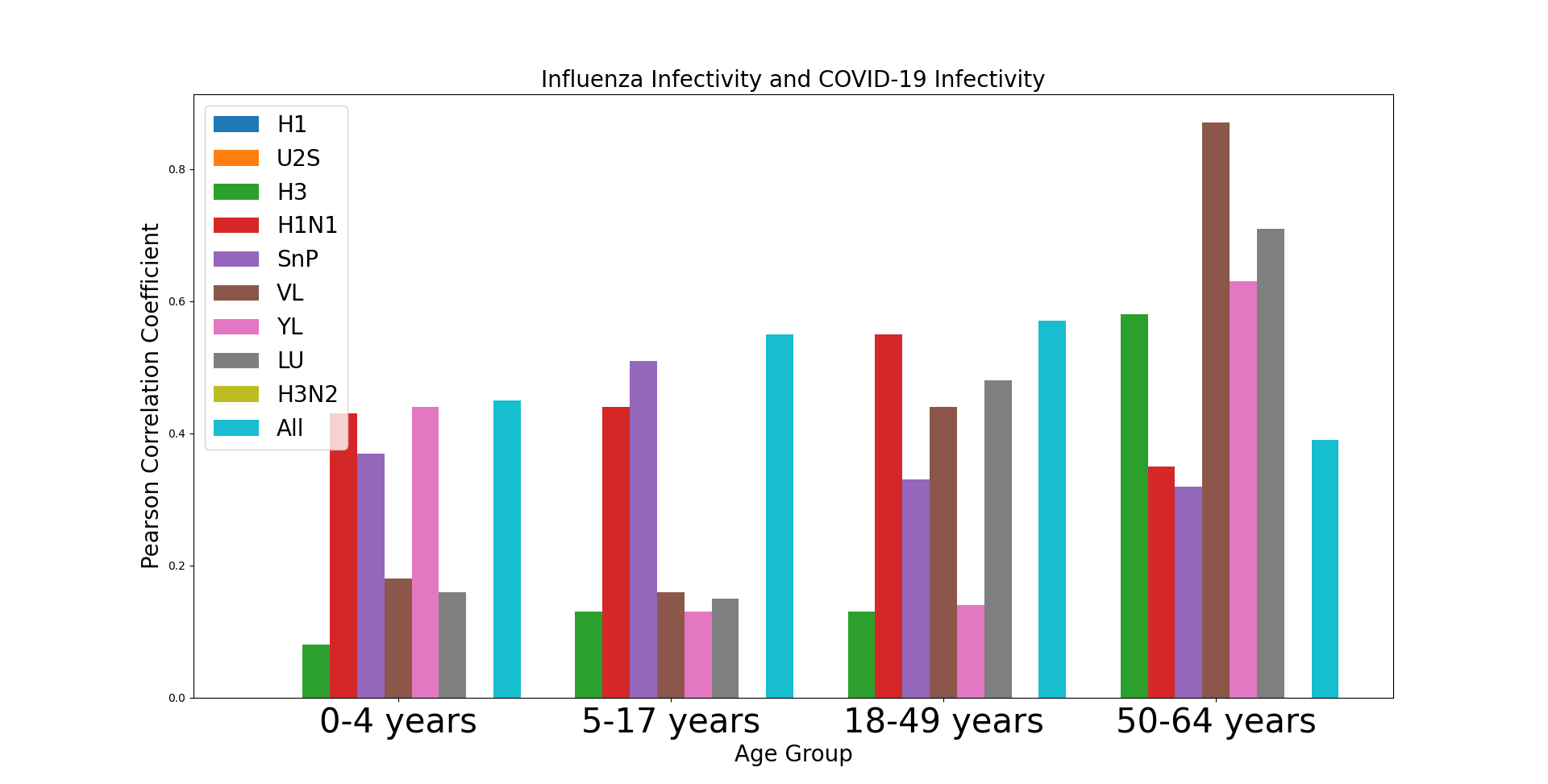


图4.10 2020年第40周到第48周流感和新冠感染力的皮尔逊相关系数

Fig.4.10 Pearson correlation coefficient of influenza and COVID-19 infection from week 40 to week 48, 2020

最终的预测结果如图4.11所示，实线表示真实的感染率，虚线是预测值；相同颜色表示相同的年龄分组。相比于2020年第10周至第20周，模型的预测效果得到了有效提升，能够比较准确地预测新冠肺炎的发展趋势。

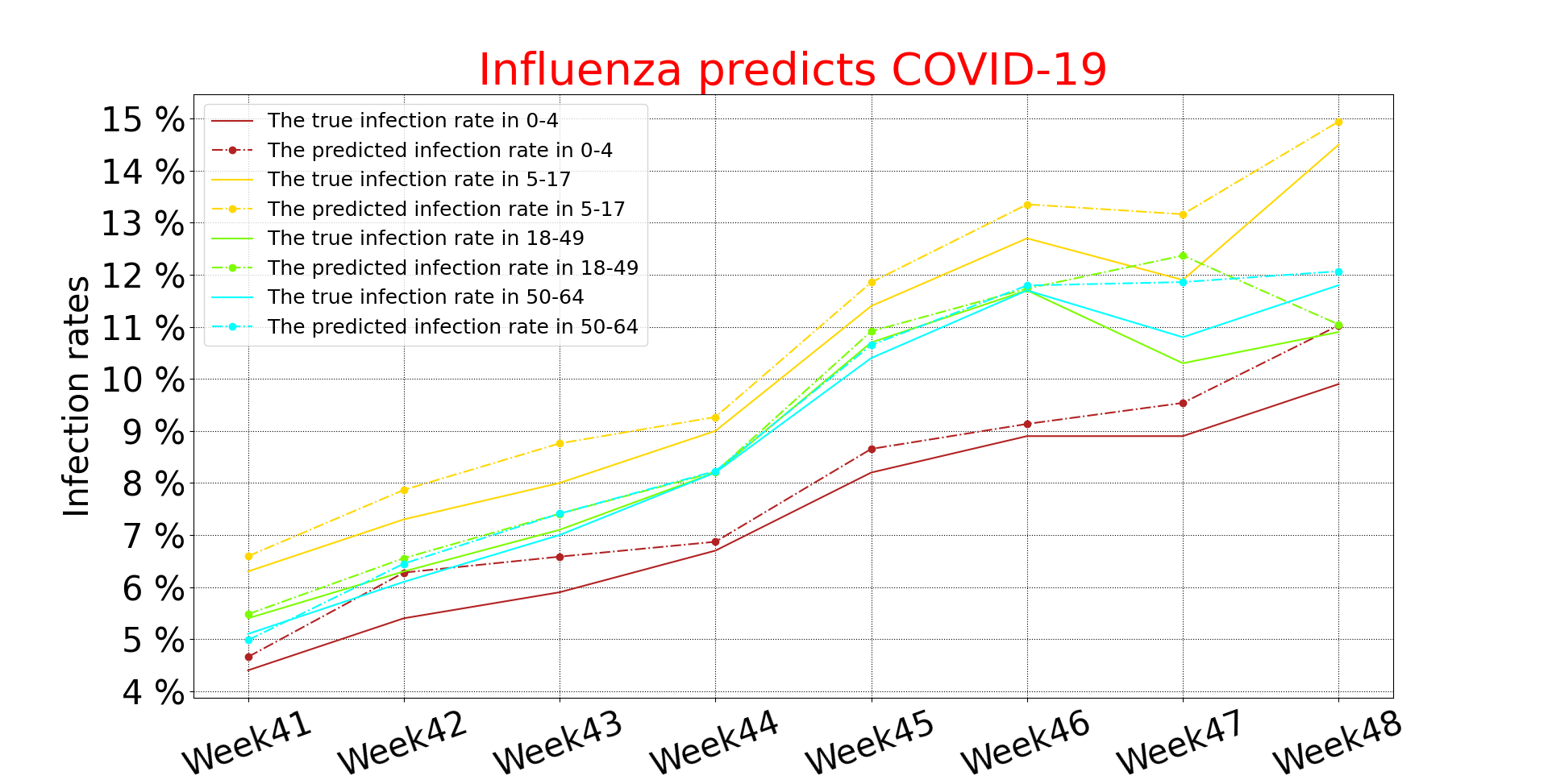


图4.11 2020年第40周到第48周流感预测新冠结果

Fig.4.11 Influenza forecast COVID-19 results from week 40 to week 48 of 2020

同样地，依然采取皮尔逊相关系数来作为衡量模型泛化性能的标准。为了验证4.3.2、4.3.3中发现的正确性，我依然用所有流感数据进行预测，相关系数柱状图如图4.12所示。

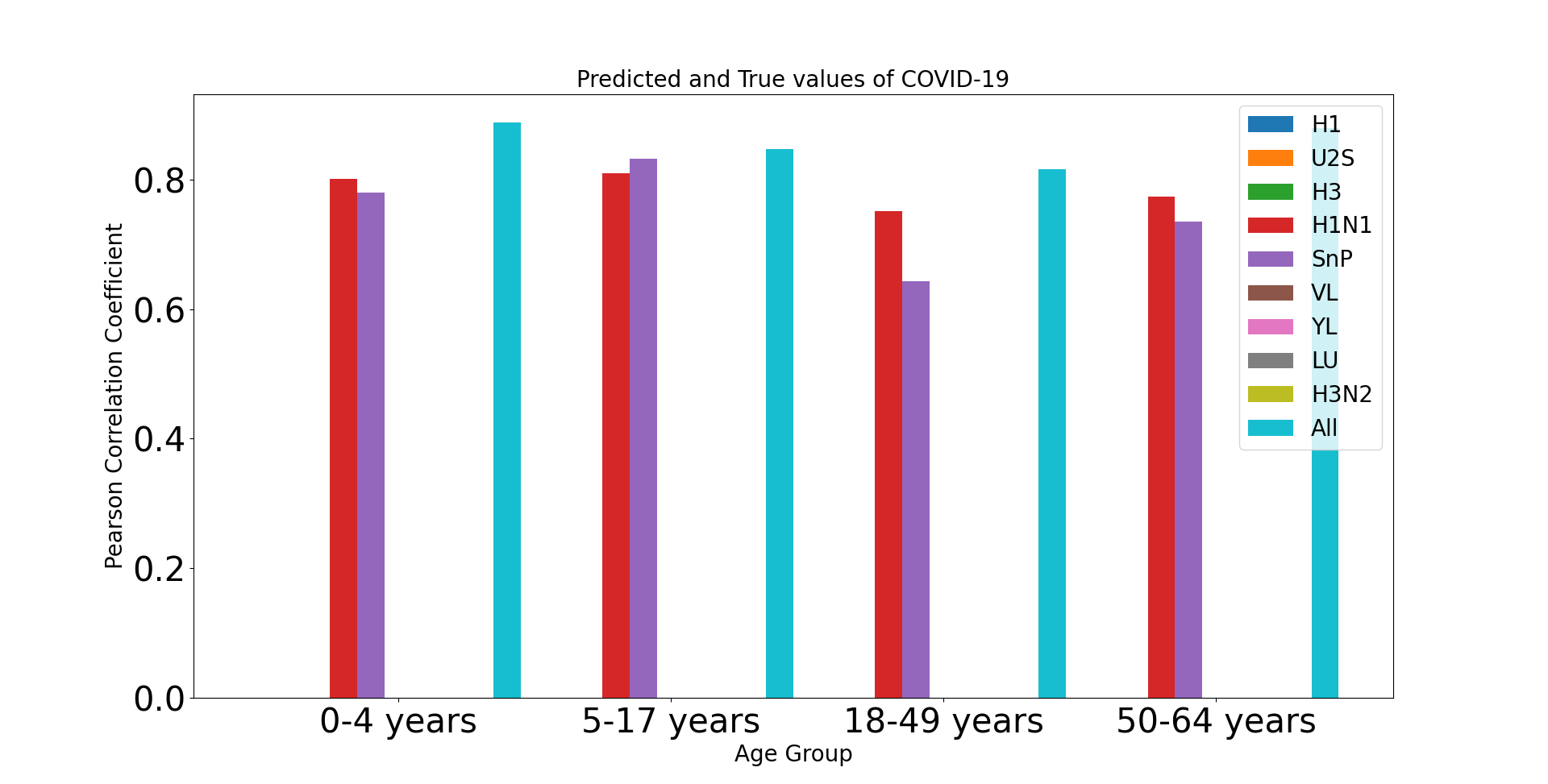


图4.12 2020年第40周到第48周预测值和真实值皮尔逊相关系数

Fig.4.12 Pearson correlation coefficient between predicted value and true value from 40 to 48 weeks in 2020

通过理论分析得到的采用All数据会生成最佳预测模型是正确的，All的相关系数在四个年龄组中均最大，且均大于0.8，预测值和真实值是强线性相关的。H1N1、SnP、All预测结果和真实值间存在线性相关，这和接触率矩阵相关系数图中的信息一致；预测相关性的排序结果与感染力相关性排序一致。至此，我们验证了4.3.2和4.3.3中发现的准确性，并得出结论：感染力相关系数决定了预测结果和真实值相关系数的大小；接触率矩阵决定了预测结果是否和真实值相关。

此外，通过将本节预测结果和2020年第10周预测结果对比，可以确定，数据的不准确性是影响本文模型泛化性能的一个主要因素。

## 4.4本章小结

本章介绍了本文实验的详细过程，本文提出模型在2020年第40-48周新冠疫情发展趋势的预测上有比较好的效果。由此，本文提出疑问：2020年10-20周统计到的疫情变化是否是真实的。在这一章中，得出了本文提出模型进行预测时的两个结论：1.感染力与待预测疾病感染力的相关系数决定了预测结果和真实值相关系数的大小；2.接触率矩阵和待预测疾病的接触率矩阵相关，则预测结果与预测疾病的真实值相关。

# 第5章 总结与展望

本文的模型可以比较准确地预测2020年第40-48周新冠疫情的发展趋势，由于流感具有季节性，且多发生在秋、冬、春季节交替之时，流感数据在夏天是有限的，且不能用来预测新冠疫情。因此本文在认为2020年第10-20周新冠数据不准确，决定更换数据时没有采用20-40周之间的数据。排除偶然性因素，本文提出的基于流感监控数据的新冠疫情发展趋势预测模型是有效的，预测结果和真实值之间的相关系数高达8。

本文旨在探索出一套新的传染病预测体系框架：根据一种已知且监测体系成熟的传染病数据，预测新流行的或未知的传染病。框架使用的前提：一是传染病通过人们的密切接触传播。本文在做实验时还探索了数据对模型泛化能力的影响情况，解决了该采用何种数据才能达到最好的预测效果这一问题。本文是在充分分析了流感和COVID-19在病理学特征、传播方式、传播模式的相似性的基础上，采用流感数据预测新冠疫情的发展趋势。但是，不能排除遇到完全未知且需要进行大量且持久的研究才能揭开面纱的传染病。在这种情况下，无法寻找与新传染病相似的疾病，因此需要探索应该采用何种传染病的数据进行预测：1.感染力线性相关；2.接触率矩阵线性相关。满足以上两个条件即可以预测。

本文在实验过程中假设各个年龄段易感人群在接触感染者后被感染的概率是相同的，应探索出不同人群接触后感染的概率，然后探索采用按年龄划分的接触感染概率后是否会提升本文模型的泛化性能。

本文在计算接触率矩阵时采用奇异值分解方法，结果不具有现实意义。应分析真实世界人群的接触情况，在求解接触率矩阵时添加限制条件，提高模型的可信度。

本文未将该框架应用到其他疾病的预测上，需要验证框架是否具有普适性。

未来工作：1.从算法的角度修改模型，提升模型泛化能力。2.进行关键参数q（接触后感染的概率）讨论，分析在不同年龄接触病原体后感染概率不同的条件下，模型的预测能力。3.查找文献或自行实验，探索真实的人群接触情况，让接触率矩阵更具有现实意义。4.验证本文提出框架能否运用到其他传染病的预测工作中。

## 参考文献

[1] Fauci A. S., Lane H. C., Redfield R. R. Covid-19 - Navigating the Uncharted[J]. ***The New England journal of medicine***, 2020, 382(13):1268-1269.

[2] 高翔, 郁建兴. 新冠肺炎疫情防控中的公共治理机制:信息,决策与执行[J]. ***治理研究***, 2020, 036(002):P.41-50.

[3] Covid T. I., Reiner R. C., Barber R. M., et al. Modeling COVID-19 scenarios for the United States[J]. ***Nature medicine***, 2020, 27(23):94-105.

[4] Covid C. D. C., Team R., COVID C. D. C., et al. Characteristics of health care personnel with COVID-19—United States, February 12–April 9, 2020[J]. ***Morbidity and Mortality Weekly Report***, 2020, 69(15): 477.

[5] Covid C. D. C., Team R., Jorden M. A., et al. Evidence for limited early spread of COVID-19 within the United States, January–February 2020[J]. ***Morbidity and Mortality Weekly Report***, 2020, 69(22): 680.

[6] Shi Y., Wang G., Cai X., et al. An overview of COVID-19[J]. ***Journal of Zhejiang University***. ***Science. B***, 2020, 21(5):343-360.

[7] 陈伟, 王晴, 李媛秋, 等. 我国新型冠状病毒肺炎疫情早期围堵策略概述[J]. ***中华预防医学杂志***, 2020, 54(3): 239-244.

[8] 唐三一, 肖燕妮, 彭志行, 等. 新型冠状病毒肺炎疫情预测建模, 数据融合与防控策略分析[J]. ***中华流行病学杂志***, 2020, 41(4): 480-484.

[9] 范如国, 王奕博, 罗明, 等. 基于 SEIR 的新冠肺炎传播模型及拐点预测分析[J]. ***电子科技大学学报,*** 2020, 49(3): 369-374.

[10] Li M. Y., Muldowney J. S. Global stability for the SEIR model in epidemiology[J]. ***Mathematical biosciences***, 1995, 125(2): 155-164.

[11] He S., Peng Y., Sun K. SEIR modeling of the COVID-19 and its dynamics[J]. ***Nonlinear Dynamics***, 2020, 101(3): 1667-1680.

[12] Kushwaha S., Bahl S., Bagha A. K., et al. Significant applications of machine learning for COVID-19 pandemic[J]. ***Journal of Industrial Integration and Management***, 2020, 5(4):361-267.

[13] Zoabi Y., Deri-Rozov S., Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms[J]. ***npj digital medicine***, 2021, 4(1): 1-5.

[14] 李忠奇, 陶必林, 占梦瑶, 等. 时间序列模型应用于新型冠状病毒肺炎疫情预测效果比较研究[J]. ***中华流行病学杂志***, 2021, 42(3): 421-426.

[15] Block P., Hoffman M., Raabe I. J., et al. Social network-based distancing strategies to flatten the COVID-19 curve in a post-lockdown world[J]. ***Nature Human Behaviour***, 2020, 4(6): 588-596.

[16] Menni C., Valdes A. M., Freidin M. B., et al. Real-time tracking of self-reported symptoms to predict potential COVID-19[J]. ***Nature medicine***, 2020, 26(7): 1037-1040.

[17] Roda W. C., Varughese M. B., Han D., et al. Why is it difficult to accurately predict the COVID-19 epidemic?[J]. ***Infectious Disease Modelling***, 2020, 5(1): 271-281.

[18] Daughton C. The international imperative to rapidly and inexpensively monitor community-wide Covid-19 infection status and trends[J]. ***The Science of the total environment***, 2020, 7(26): 138149.

[19] Sun S., Folarin A. A., Ranjan Y., et al. Using Smartphones and Wearable Devices to Monitor Behavioral Changes During COVID-19[J]. ***Journal of medical Internet research***, 2020, 22(9): 19992.

[20] Liu Y., Mao B., Liang S., et al. Association between age and clinical characteristics and outcomes of COVID-19[J]. ***European Respiratory Journal***, 2020, 55(5):321-333.

[21] Hutchinson E. C. Influenza virus[J]. ***Trends in microbiology***, 2018, 26(9): 809-810.

[22] Cao X. COVID-19: immunopathology and its implications for therapy[J]. ***Nature reviews immunology***, 2020, 20(5): 269-270.

[23] Solomon I. H., Normandin E., Bhattacharyya S., et al. Neuropathological features of Covid-19[J]. ***New England Journal of Medicine***, 2020, 383(10): 989-992.

[24] Rolfes M. A., Foppa I. M., Garg S., et al. Annual estimates of the burden of seasonal influenza in the United States: a tool for strengthening influenza surveillance and preparedness[J]. ***Influenza and other respiratory viruses***, 2018, 12(1): 132-137.

[25] 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.

[26] Carcione J. M., Santos J. E., Bagaini C., et al. A simulation of a COVID-19 epidemic based on a deterministic SEIR model[J]. ***Frontiers in public health***, 2020, 8(3): 230.

[27] Edmunds W. J., Gay N. J., Kretzschmar M., et al. The pre-vaccination epidemiology of measles, mumps and rubella in Europe: implications for modelling studies[J]. ***Epidemiology & Infection***, 2000, 125(3): 635-650.

[28] Rohani P., Zhong X., King A. A. Contact network structure explains the changing epidemiology of pertussis[J]. ***Science***, 2010, 330(6006): 982-985.

[29] Torres A., Sligl W., Lee N. Severe flu management: a point of view[J]. ***Intensive care medicine***, 2020, 46(2): 153-162.

[30] Edridge A. W. D., Kaczorowska J., Hoste A. C. R., et al. Seasonal coronavirus protective immunity is short-lasting[J]. ***Nature medicine***, 2020, 26(11): 1691-1693.

## 致 谢

感谢我的论文指导老师陶丽老师，感谢老师每周能在百忙之中抽出时间与我讨论，在实验过程以及论文撰写时对于我所遇到的困难和疑惑给出了悉心的指导。同时感谢老师关怀我的身体状况，有一段时间我每晚失眠，白天昏昏欲睡却又硬撑不睡，导致体重骤降，身体素质变差。陶老师看到我的状况后，十分关心，让我倍感温暖。正是陶老师的关心，让我重整状态，学习和生活再次步入正轨。

感谢父母对我生活和学习上的支持。无论我提出什么条件，只要是和学习相关，父母都会毫不犹豫地答应我。

感谢巍巍学府西南大学给我提供学习的芝兰之室。身为河北考生，我深刻体会到了想要考入一所优秀的高校是何其困难。当时高成考成绩并不理想，感谢西南大学当时录取我，让我能在如此好的环境中学习。我在西南大学度过了最美好的大学时光，在这里我遇到了和蔼可亲、教导有方的老师；遇到了志同道合、朝夕相处的朋友。

最后感谢论文评阅老师的辛苦工作，感谢参加答辩的各位老师。