分类号 C8/283　　密级 公开

U D C　　编号 10741

兰州财经大学

LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

（专业学位）

论文题目 全国乙肝流行特征分析及发病预测研究

研 究 生 姓 名：毛少霞

指导教师姓名、职称：赵煜 副教授

学 科、专 业 名 称：统计学 应用统计硕士

研 究 方 向：大数据分析

提 交 日 期：2021 年 6 月 6 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：毛少霞 签字日期：2021.6.6

导师签名：赵 华 签字日期：2021.6.6

导师(校外)签名：曹正凤 签字日期：2021.6.6

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：毛少霞 签字日期：2021.6.6

导师签名：赵 华 签字日期：2021.6.6

导师(校外)签名：曹正凤 签字日期：2021.6.6

Analysis of epidemic characteristics and prediction of Hepatitis-B in China

Candidate : Mao Shaoxia

Supervisor: Zhao Yu

摘 要

传染病不仅影响人类健康，而且对地区经济发展产生约束。国家疾病预防控制局报告显示，乙肝发病数在传染病例总数中占有很高比重。由于当前发病人数较多，能接受规范性抗病毒治疗的比例较低，防治乙肝任重道远，了解我国乙肝的流行特征并进行合理预测是防治乙肝的关键。

论文以我国乙肝发病率为研究对象，选取 2004-2017 年乙肝发病率的时空多维度数据，就乙肝发病率基本特征及预测方法进行挖掘与探讨。首先利用传统流行病学中的“三间分布”对不同维度的乙肝发病情况进行描述性分析，探讨乙肝的分布差异；其次结合空间统计方法和 GIS 软件挖掘其空间特征，一方面，从可视化角度制作我国乙肝疾病专题地图，以直观展示乙肝发病率的空间分布特征。另一方面，借助 Moran’s I 指数和 LISA 集聚图，进一步捕捉其空间集聚特征；最后探讨乙肝发病率的科学预测方法，通过比较单一模型和组合模型，最终选取合理模型对其未来趋势进行预测。

根据上述分析得出以下结果：

（1）乙肝在不同时间、不同地区和不同人群间的分布都存在明显差异。时间分布中，不同年份的发病率波动范围在 68.57/10 万到 89.00/10 万之间，2009 年发病数最多达到 1179607 例，3 月份是乙肝的高发病期，12 月份乙肝发病数最少。地区分布中，不同地区的乙肝年发病率在 7.79/10 万-251.05/10 万之间，2004-2017 年广东省发病数最多，西藏发病数最少。性别层面，男性发病率明显高于女性发病率。年龄层面，0-19 岁间的低年龄段发病数逐渐减少，40 岁以上人群的发病数逐渐增多。职业层面，农民乙肝发病数最高，其次是家政、家务及待业和工人。

（2）空间相关性分析结果表明，2004-2017 年各省市区乙肝发病率存在较强的空间正相关性。发病率具有明显的空间聚集效应，主要集聚特征为高-高发病率地区相邻、低-低发病率地区相邻，其中高发病率地区主要位于西部的青海、新疆和南部的广东，低发病率地区主要包含上海和江苏。

（3）使用三种模型预测乙肝月发病率，结果显示 ARIMA-BP 神经网络组合模型预测效果最优，该模型各项评价指标得出的误差都较小，模型拟合精度高，最终使用该组合模型预测乙肝发病率的未来趋势。

关键词：乙肝流行特征　预测　三间分布　Moran’s I 指数　ARIMA-BP 组合模型

Abstract

Infectious diseases affect human health, it can also hinder regional economic development. According to the report of Chinese Centre for Disease Control and prevention, the number of cases of Hepatitis-B accounted for a high proportion of the total number of infectious cases. Because of the large number of patients, the proportion of patients who can receive standardized antiviral treatment is low. It is an urgent issue to prevent, diagnosis and treatment of Hepatitis-B. An important issue for prevention and treatment of Hepatitis-B is to understand the epidemic characteristics and introduce a reasonable prediction.

Taking the incidence of Hepatitis-B in China as the research object, this paper selected the spatial and temporal multi-dimensional data of the incidence of Hepatitis-B from 2004 to 2017 to dig and explore the basic characteristics and prediction methods of the incidence of Hepatitis-B. Firstly, a descriptive analysis of the incidence of Hepatitis-B in different dimensions was carried out by using the "three-space distribution" in traditional epidemiology to explore the distribution differences of Hepatitis-B. Secondly, spatial statistical methods and GIS software are combined to mine its spatial characteristics. On the one hand, a thematic map of Hepatitis-B disease in China is made from the perspective of visualization to visually display the spatial distribution characteristics of the incidence of Hepatitis-B. On the other hand, Moran's index and LISA

cluster map were used to further capture the spatial clustering characteristics. Finally, the scientific prediction method of the incidence of Hepatitis-B was discussed. By comparing single model and combination model, a reasonable model was selected to predict the future trend of Hepatitis-B.

The following results are obtained.

(1) There are significant differences in the distribution of Hepatitis-B at different times, different population and different regions. In the time distribution, the incidence of Hepatitis-B in different years fluctuated from 68.57/100,000 to 89.00/100,000. In 2009, the number of cases reached the maximum of 1179,607 cases. March was the high incidence period of Hepatitis-B, and December was the lowest. In terms of regional distribution, the annual incidence of Hepatitis-B in different regions ranged from 7.79/100,000 to 251.05/100,000. From 2004 to 2017, Guangdong Province had the highest incidence, while Xizang Province had the lowest incidence. The incidence rate of male is significantly higher than that of female. The incidence of low age among 0-19 is gradually decreasing, and the incidence rate of people aged 40 or above is increasing. The highest incidence was found in farmers, followed by housekeeping, housework, unemployed and workers.

(2) Spatial correlation analysis showed that there was a strong spatial positive correlation between the incidence rates of Hepatitis-B in each

province from 2004 to 2017. The incidence rate is obviously spatial aggregation effect, and the main clustering characteristics are high- high incidence rate adjacent area and low-low incidence area. High incidence rate is mainly located in Qinghai, Xinjiang and Guangdong in the west, and the low incidence rate mainly includes Shanghai and Jiangsu.

(3) Fitting the prediction of the incidence rate of Hepatitis-B through three models, the results show that the ARIMA-BPNN combined model has the best prediction effect, and the error of each evaluation index of this model is small. The model has high precision and can be used for the prediction of Hepatitis-B incidence rate.

Keywords: Epidemic characteristics of Hepatitis-B; Forecast; Three distribution; Moran’s Index; ARIMA-BPNN hybrid model

目 录

1 引 言

1.1 研究背景

传染病可以在不同生物间传播并迅速蔓延，对人类健康、区域经济发展方面都产生负面影响。虽然世界各国为预防和防治传染病已经做出各种努力，但控制其蔓延，减少其危害仍然是人类社会面临的重大挑战。病毒性肝炎由于其对患者带来的痛苦和严重的疾病负担，是传染病控制中很重要的一个内容，而其中的乙型病毒性肝炎作为乙类传染病，虽然我国已经有一整套以“预防为主，防治结合”的防控体系，但由于患者数量大、诊断率和治疗率较低，还需进一步构建科学有效的预防及监管机制。

根据乙肝流行的严重程度，世界卫生组织通过 HBsAg 阳性率将不同区域分为高、中、低三个流行区，相对应的，三种流行区的 HBsAg 阳性率分别为 8%-20%、2%-7%、0.2%-0.5%[21]。2019 年“世界肝炎日”报道显示，全球乙肝流行率最高的地区是西太平洋和非洲地区，成人感染比例分别达到了 6.2%和 6.1%。中国人口数占西太平洋区域人口数的 78%，这也表明我国乙肝成人感染比例较高。

从血源性乙肝疫苗在我国的最初使用，到目前全面开展乙肝疫苗免费接种政策，这一系列防治措施取得了一定成效，乙肝感染率明显下降，我国已经不再属于高流行区[7]。此外，我国加强乙肝诊疗监管和筛查，降低母婴和血液传播的风险，也在一定程度减少了乙肝发病数。然而，根据国家疾控部门公布的数据显示，2018 年和 2019 年我国病毒性肝炎人数在所有传染病人数中占比分别为 41.78%和 41.88%，2019 年肝炎发病数比重反增不减。目前的统计资料显示，截止 2020 年 3 月，病毒性肝炎发病数仍然很高，死亡率排名第三，其中乙型肝炎的发病数为 88150，死亡数 31，乙肝发病数在甲乙丙类传染病中占比达到最高。

乙肝的发展速度很快，病毒感染使患者肝脏功能异常。患者不但要忍受疾病带来的巨大痛苦，而且要承担严重疾病负担。尽管乙肝预防政策已经取得很大成就，降低了乙肝发病率，但由于患者人群基数大，患者不能及时接受规范性抗病毒治疗，乙肝患者防治迫在眉睫，肝炎的防治任务依然很严峻。所以根据乙肝现状进行科学防治，科学监测，在预测发病率基础上及时布局是乙肝防治的重要环

节。

1.2 研究目的及意义

现有的疾病分布特征研究方法偏向于传统描述性统计，乙肝发病率预测模型的精度较低，不便于乙肝防治中的科学决策，所以在后续乙肝防治工作中，如何构建科学合理的量化分析体系和及时反馈的动态监管体系是需要进一步做的工作，而其中科学有效地了解疾病分布特征和疾病预测方法的探讨是值得研究的问题。

论文结合传统流行病学方法和先进的统计分析方法，对乙肝流行病学特征及未来趋势做深入研究。从传统流行病学角度分析乙肝发病数及发病率的地区差异、时间差异和人群差异，全面了解我国乙肝发病率的分布情况。此外，利用 GIS 地理信息系统对乙肝疫情的空间分布模式和集聚情况进行探索性分析，充分了解我国乙肝流行特征，找出乙肝的高发地区和高危人群，以便做好乙肝免疫策略调整，有针对性地防控乙肝和降低乙肝发病率。

在了解疾病流行特征的基础上，构建传染病预测模型为及时布局防控乙肝奠定了理论基础。根据现有的乙肝月发病率数据，利用不同模型进行拟合并选择最优模型展开预测。准确的传染病预测模型可以掌握疾病流行发展的过程和规律，为流行病的科学监测和防治提供参考依据。根据研究结果，相关部门可以针对发病率的“热点”区域和“冷点”区域，制定合理有效地防控措施和监管体系。此外，论文构建的传染病预测模型，也可以尝试使用在其他疾病预测中，为后续传染病预测提供新的思路。

1.3 国内外研究进展

1.3.1 流行病学方法研究进展

（1）传统流行病特征分析

理论上讲，流行病学研究的最终目的是提出对策、采取措施控制和消灭疾病，整个研究过程的关键在于发现疾病分布规律，找到引起疾病的原因。在实际分析中，主要通过调查了解疾病发病现状，探讨疾病在不同维度（地区、时间、人群）

发病情况和死亡情况。国内外众多学者从不同角度探讨疾病的流行病学特征。从研究方法来看，问卷调查方法和描述性统计方法已经非常成熟，都被广泛用于流行病学研究，一些学者通过调查问卷和抽取血液标本对乙肝相关抗原抗体检测，绘制统计图表刻画乙肝在不同人群中的分布，发现乙肝传染的内在规律；还有部分学者发放问卷调查乙肝发病人群或公众（医护人员）对乙肝的认知情况，根据被调查者具体行为评估不同人群对乙肝相关知识的了解程度。

例如，史雯[27]等以浙江省健康人群为研究对象，对其展开乙型肝炎血清流行病学调查，调查结果显示：乙肝接种率随年龄增长呈逐渐下降趋势，乙肝表面抗原（HBsAg）阳性率男性高于女性，乙肝疫苗接种史对成人 HBsAg 阳性率有很大的影响；邓秋云[16]调查 1-59 岁人群，发现年龄、性别、乙肝疫苗接种史、地区、民族、职业、文化程度、婚姻状况影响 HBsAg 阳性率；崔富强[15]等分析了 2005-2010 年我国乙型病毒性肝炎发病情况得出，15 岁以下的病例数占比有明显下降趋势，已经从 5.56%下降为 1.92%。范珂[17]等调查了 300 多名患者对乙肝防治知识的了解程度，其中包括对乙肝的态度和乙肝相关行为的调查，结果显示乙肝患者对乙肝的认知不够全面，年龄、学历以及职业影响患者对乙肝的了解程度；刘观秀[23]以普通大众为调查对象，从乙肝疫苗知识、接种疫苗态度及疫苗接种情况三个角度展开分析，调查结果表明，大众对乙肝疫苗的知信行情况较好，但仍需加强对男性和低学历人群乙肝疫苗相关知识的宣传教育。

（2）空间流行病分布特征分析

早在 1854 年，国外学者就将空间分析技术用于流行病研究，当时伦敦爆发疫情，英国医生约翰·斯诺[3]首次在伦敦地图标出当地霍乱的发病情况，通过分析霍乱患者居住位置而确定引发霍乱的原因，最终有效控制了霍乱的传播。随着计算机技术的迅速发展，地理信息技术在空间流行病学有了更深入的应用，学者们进一步分析疾病的空间分布特征，结合研究对象的地理位置提取流行病发病的空间信息成为空间流行病学不断探索的内容。EmmaK 等[1]通过 GIS 地图，根据空间聚类方法得出研究地区的疾病分布不是随机分布，进一步找到该疾病的高发地区。Yuliang Xi [11]等对 2010 年深圳市住院的 2851 例乙肝病例进行了研究，结果显示乙肝空间分布与风险因素、区域医疗资源空间获取不均，主要集中在深圳南部和西南部。我国学者瞿嵘[38]以深圳为研究区域，利用空间数据分析技术探讨

疾病分布差异和变化规律，通过空间扫描统计方法得出乙肝在时空上存在聚集特征。覃柳麻[26]分析南宁市 2012-2017 年乙肝分布情况，同样得出相似结果，即南宁市乙肝在空间分布上存在空间聚集性。肖占沛[32]分析河南 2009-2019 年风疹的空间相关性，结果显示河南省风疹发病存在时空聚集性，其风疹发病呈显著下降趋势。众多学者将 GIS[44]及空间数据分析技术应用到流行病学中，通过挖掘疾病的空间信息，探究出流行病存在空间集聚效应。这些方法的使用推动了空间流行病学的进一步发展，为流行病防控和及时布局做出了很大贡献。综上所述，以空间视角探索流行病发病规律，考虑地理环境等因素对疾病的影响，便于全面了解疾病流行特征和传播现状，以期对未来疾病的防控及量化研究奠定理论基础。

1.3.2 传染病预测方法研究进展

预测的目的就是通过研究历史数据中存在的规律和特征，分析下一阶段的变化趋势。由于传染病具有发展迅速，传播较快等特点，构建合理有效的发病率预测模型可以为疾病防控提供理论依据，促使相关部门在防控时及时采取措施，在应对问题时更加有方向感和着力点，减少控制疾病所需的时间和费用。计算机技术的发展推动传染病预测方法日趋完善，目前用于传染病发病率拟合和预测的方法很多，其中包括传统的时间序列分析，空间计量方法以及机器学习算法等，时间序列分析从发病数据本身出发，根据数据特性构建合理模型；空间计量方法更关注影响疾病的地理因素和空间结构；机器学习算法可以不断训练，对非线性趋势拟合效果优良。这些方法在疾病预测中发挥重要作用。以下主要将传染病预测方法分为单一模型和组合模型展开综述。

（1）单一预测模型

众多学者将时间序列预测方法用于传染病发病研究中，其中 ARIMA 模型被广泛应用在实际研究中，陈婷[14]用 ARIMA 模型对 2015 年我国艾滋病发病率进行预测，该模型预测结果和实际数据差别小，短期内的拟合效果较好。王平[30]通过对比指数曲线模型、灰色系统模型和 ARIMA 模型在病毒性肝炎、痢疾和麻疹预测中的应用，最终发现前两种疾病预测中 ARIMA 模型最优，而对麻疹发病率进行预测时，三种模型均无效。杨晓丽[36]用季节乘积型 ARIMA 模型预测辽宁省乙肝发病趋势，预测结果显示该模型平均相对误差为 7.12%，可用于短期预测。

以上研究表明，用单一时间序列方法进行传染病预测时，ARIMA 模型的适用性更强，预测效果也较好。

随着机器学习算法的流行，BP 神经网络模型、支持向量机等方法在传染病预测中发挥着越来越重要的作用。陈婷[14]预测艾滋病月发病率发现：BP 神经网络模型在艾滋病月发病率预测中比 ARIMA 模型更占优。于颖慧[37]考虑周期性影响，用 SARIMA、支持向量回归机、小波神经网络、极端学习机和非线性自回归滤波四种模型预测手足病发病率并评估预测性能，结果表明 ELM 模型的预测效果优于其他四种模型。这些预测模型都是考虑发病序列本身的变化规律，进而展开了预测研究。

此外，应用灰色模型对原始数据进行处理，也可以发掘数据间的变动规律并预测事物未来发展情况，由于该模型对原始数据约束较少，因此也有学者将灰色预测模型应用到传染病预测中，最终取得不错的预测效果。周强[42]等在徐州市乙肝发病趋势研究中使用灰色系统 GM(1,1)模型，预测得出徐州市未来乙肝发病率逐渐下降，张靳冬[40]等同样用该灰色模型预测常州市乙肝发病趋势，结果显示 GM(1,1)模型有较好的拟合效果。Ya-wen Wang[10]等对比 ARIMA 模型和 GM(1,1)模型对乙肝发病的预测效果，其中 ARIMA 模型精度高，预测效果更好。除了上述几种预测方法，马尔可夫模型也被用于疾病预测中，杨品超[35]等从乙肝防治的经济学角度出发，构建了马尔科夫模型，并通过验证说明该模型与我国实际情况相符，适用于我国乙肝防治策略评价中。Shahdoust Maryam[8]等采用基于马尔可夫链理论的加权马尔可夫链（WMC）方法、Holt 指数平滑（HES）以及 SARIMA 两种时间序列模型对乙肝数据进行处理，结果证明 HES 模型对发病率的预测最为准确。刘琼[24]等建立的单变量正态分布隐马尔科夫模型，对于预测未来乙肝疫情有其应用价值。

（2） 组合预测模型

在传统时间序列方法和机器学习方法基础上，利用单一预测方法构建组合模型为疾病预测提供了新思路。付之鸥[17]等考虑时间序列包含的非线性趋势，构造了 ARIMA-SVR 和 ARIMA-BPANN 组合模型，相比单一线性预测模型和非线性预测模型，这些组合模型拟合效果最优。谢骁旭[33]对 ARIMA 残差序列进行重构，建立 ARIMA-SVM 组合模型，分别使用单一模型和组合模型预测江西省肺

结核发病率，根据选取的两个评价指标（MSE 和 MAPE）得出，ARIMA-SVM 组合预测预测结果更优。乔贺倩[25]用小波分析和奇异谱分析分解时间序列，针对不同分解项选取合适模型预测发病情况，一共得到 4 个组合模型，原始序列和组合模型分别建模比较预测准确度，组合模型更充分提取了序列信息，预测效果更好。总的来看，大部分组合模型综合了两种基础模型的优势，在预测发病情况时优于单一模型，预测效果更佳。

1.3.3 研究述评

通过了解国内外文献可知，传染病流行特征分析和发病预测方法取得很大成果，但仍存在一些不足。

第一，已有的疾病流行特征分析主要围绕传统描述性统计，其强调流行病学研究中的观察分析，将主要工作内容放在统计和描述疾病发病率、患病率等指标。该描述性统计方法对发病研究起到了很重要的作用，但却忽略了疾病的动态发展和地理环境的影响。虽然地统计学方法有一定的介入，但有关空间统计的挖掘不够深入。为此，论文采用时空结合的分析方法，探讨我国乙肝发病数据的规律，从整体角度全面把握全国乙肝的流行特征，通过不同维度深入挖掘乙肝发病信息，为后期国家准确、有针对性地实施乙肝防控措施提供依据。

第二，已有的传染病因果预测模型较少，由于影响传染病的主要因素不易找，且人口流动性大，该预测方法在具体实施中较困难，而传统时间序列模型在传染病预测中表现出其优良性能，不仅模型构造简单，短期预测效果也较好，因此被众多研究者广泛使用，但该模型不能充分提取非线性特征，模型长期拟合效果不佳。此外，有部分学者考虑传染病发病率序列的波动情况，利用机器学习方法对有波动的非线性趋势进行预测。这些单一模型对发病信息提取不够全面。所以，综合考虑单一预测模型优势，通过组合预测模型充分挖掘发病序列中的线性趋势和非线性趋势，提升模型预测性能。

1.4 研究内容、方法与创新点

1.4.1 研究内容

在了解国内外学者对传染病相关研究的基础上，论文通过传统流行病学中的“三间分布”和空间分析技术，分析我国乙肝发病现状，挖掘不同维度下乙肝分布差异和流行规律。此外，在已有的乙肝发病率数据基础上，建立合适的预测模型拟合发病情况。具体研究内容如下：

第一部分，引言。这部分主要从研究背景及研究意义出发，在学习了解国内外文献的基础上，综述学者对乙肝研究概况，分析关于乙肝发病情况的研究，得出论文思路和结构。

第二部分，介绍研究所用的相关理论和方法。系统阐述理论基础和基本方法，其中包括流行病学统计方法分析、空间分析方法和传染病预测模型。

第三部分，根据已有数据，对全国乙肝发病率进行流行病学特征分析。首先通过描述性统计对乙肝在不同时间、性别、年龄，职业及地区的分布情况进行定量分析，总结出发病率随时间等变化的规律。其次，从空间角度探索乙肝在时空上的变化规律，通过空间相关性分析判断乙肝是否集聚，找到乙肝高发病区域。

第四部分，通过研究大量文献，在已有的预测模型中，基于获取的数据特征选择效果较好的时间序列模型进行预测。此外，将机器学习方法应用于预测研究，运用组合模型对发病率进行预测，找出最优的预测模型，最终对乙肝发病率进行预测，通过对比分析预测结果与实际发病数据，评价预测效果。

第五部分，结论、思考及展望。根据上述分析得出乙肝发病特征和预测相关的结论，对获得的结果展开讨论，从而发现研究内容和方法中的不足和局限，以便进一步探讨后续研究方向和需改进的地方。

1.4.2 研究方法及路线

（1）多学科交叉数据分析方法。立足多学科领域，综合使用流行病学、空间地理和统计学等相关方法，利用地理信息系统 GIS，对全国各地区乙肝发病率数据进行描述性分析、空间分布描述，从多个角度对乙肝流行特征展开分析。

（2）时空统计分析方法。通过时间维度和空间维度挖掘疾病特征，应用 ArcGIS 和 Rstudio 等软件分析不同时间疾病聚集性、地理环境与疾病相关性。

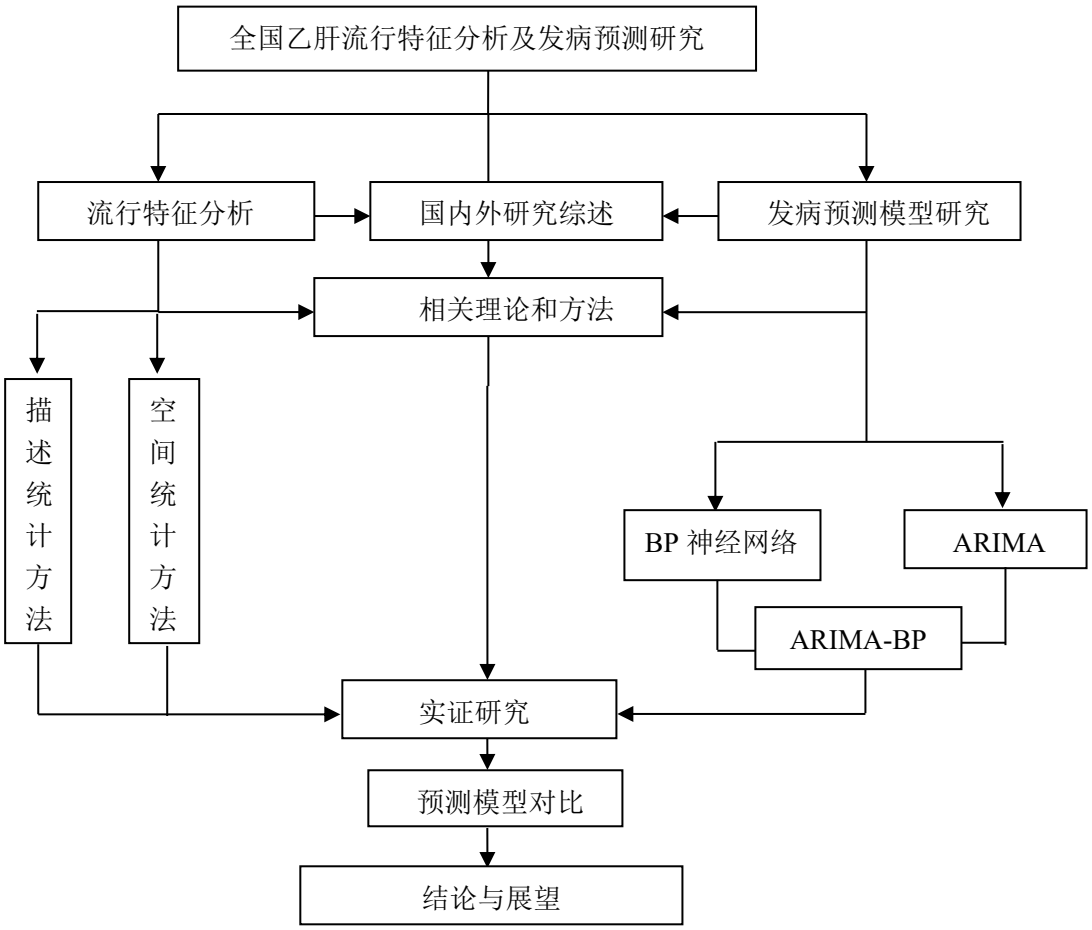


图 1.1 技术路线图

1.4.3 可能的创新点

（1）从研究范围看，大多数学者对乙肝发病率的研究都是从省市展开，考虑到随着经济发展，人口流动性强度比较大，乙肝发病率存在区域集聚特征，相比于省域和地市州层面，从全国尺度进行研究，可以更加全面了解乙肝在我国的整体流行特征，挖掘出不同省市区间乙肝集聚情况，分析出乙肝的“热点”地区和“冷点”地区，这也为具体的防范措施找到了重点和难点，因地制宜，便于相关部门针对性地采取预防措施、及时布局。

（2）采用时空相结合方法，同时考虑时间维度和空间地理维度，综合分析疾病分布情况，更吻合疾病的传染特征，便于更深层次地挖掘乙肝分布特征。

（3）从研究方法看，相比较单一预测模型，采用时间序列模型和机器学习的组合预测方法，既提取了发病序列中的季节特征，又提取了非线性特征。这种组合模式在传染病研究中有其应用价值。

2 相关理论及方法

2.1 流行病学统计相关方法

2.1.1 流行病分布特征统计

流行病学的研究和时空密不可分，传统流行病学主要通过“三间分布”[21]探讨疾病流行特征，观察疾病在人群中的发生及多少情况，描述疾病在不同维度下分布差异，根据统计分析结果反映疾病的发病规律。流行病学特征主要包括以下分布特征：

（1） 时间分布特征

疾病的发病情况在不同时期存在差异，尤其对于传染病而言，季节、环境、气候等条件的不同可能导致发病数有明显的变化，所以研究传染病的时间分布特征变得尤为重要。根据所研究时间段乙肝疾病的年(或月份)发病数及发病率等指标，分析流行病是否存在短期或长期趋势等变化规律，最终寻找发病的高峰期，为后续疾病防控提供理论依据。

（2）地区分布特征

不同疾病的分布情况各不相同，一方面，地区的饮食习惯和风土人情可能会引起一些地方性疾病的发生，导致某些疾病仅仅分布在特定地区；另一方面，传染病在全世界范围内都有分布，但由于地区地理位置及环境等其他因素的影响，地区间的发病数量和发展动态存在差异。此外，传染病在一定条件下可向周围地区波及，由于不同地区生活习惯，环境卫生条件等因素不同，传染病最终的分布情况也各不相同，分析疾病在不同地区分布差异便于清楚了解疾病发病的“热点”区域和“冷点”区域。

（3）人群分布特征

人口学分布特征主要包含三个方面：发病人群的性别、年龄和职业，由于不同人群活动范围和易感程度存在差异，最终导致疾病的分布情况有所区别。结合我国乙肝发病的具体数据，计算出男女发病数、不同年龄组发病数和不同职业发病数的构成情况，进一步得出高发病率存在于哪些年龄组和职业中，通过职业反映不同劳动者经济条件和卫生状况等因素对乙肝发病的影响，对乙肝疾病分布有

更加全面的认识。

2.1.2 流行病空间关联性统计

空间流行病学是在传统流行病学研究的基础上形成的，传统流行病学研究方法更多以特定临床观察、诊断指标等描述疾病的分布，随着测量技术和计算机技术的高速发展，空间分析技术进一步推动空间流行病学的形成和发展，不同地区的流行病数据与位置信息相结合，形成了“空间数据”。这促使我们用更加全面的角度研究流行病学的流行特征和发展趋势，区别于传统的“三间分布”，更多地从空间维度探究流行病的属性及特征。空间数据分析的主要目的是理解空间数据之间的相关性和统计关系[41]，相比传统“三间分布”，考虑疾病发病的地理位置展开分析，更符合流行病传播特点。通过该方法探讨疾病的发病特征和集聚情况，更有利于后续流行病防控和及时监管。

（1）空间权重矩阵

通过度量各区域间的邻近关系，最终确定空间权重矩阵。也就是说，属性值的相似度或相异度是根据空间中相应地理位置的距离来评估，由此确定空间目标的相对位置关系。针对不同研究内容，度量空间目标邻近性的方法也各不相同，空间权重矩阵W的具体表达形式如下：

$$W=\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix}$$

其中W_{ij}表示区域i与j的邻近关系。在实际应用中，可以分为邻接矩阵和距离矩阵。如果用简单的邻接关系来判断，则当 i与j两个位置邻接，此时空间权重矩阵中$W_{ij}=1$，相反$W_{ij}=0$。常用的 Bishop 邻接、Rook 邻接和 Queen 邻接是三种不同的相邻关系，Bishop 邻接表示该地区与其他地区是共顶点连接，Rook 邻接表示该地区与其他地区是共邻边连接，Queen 邻接相当于前两种邻接的结合，此时不仅是共顶点连接，而且是共邻边连接。

除了用相邻关系决定空间权重矩阵，另一方面还可以根据距离来描述空间单元的关系，此时不再考虑多边形之间的邻接性，而是把多边形区域表示为点，点

可以确定为多边形的质心或行政中心，然后根据点的坐标测度点之间的距离，距离远近决定了空间权重系数的大小。此外，当存在更为复杂的相邻单元对其他非相邻单元产生影响时，可根据实际情况设定其他高阶邻接关系。

空间自相关用来检验空间邻域中属性值之间是否存在关联，通过空间自相关分析技术将空间单元属性的聚集程度表现出来，简单来说，就是找某一空间单元是否和周围其他空间单元在某种属性上存在相关性。

（2） 全局空间自相关

全局空间自相关最常用的统计量是 Moran's I 指数[5]，该指标反映邻近地区间是否存在空间正相关或空间负相关关系。当 Moran's I 值为正数时，相似属性值存在聚集状态，Moran's I 为负数时，相异属性值存在聚集状态，Moran's I=0，说明高值和低值完全随机分布。Moran's I 指数的具体计算公式如 2-1 所示：

$$I=\frac{n\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}(x_i-\overline{x})(x_j-\overline{x})}{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}(x_i-\overline{x})^2} \quad (2\text{-}1)$$

上式中，n 代表所研究对象的个数，在该研究中表示我国 31 个省市区数目，x_i, y_j 表示不同地区的属性值，$\overline{x}$ 在论文中指全国乙肝平均发病率，W_{ij} 代表空间权重矩阵，可根据各省市之间的邻近程度来确定。

对 Moran's I 指数进行假设检验，建立原假设：研究单元之间无空间相关性。在给定显著性水平 0.05 的条件下，判断是否拒绝原假设[28]。当 $|Z|>1.96$ 时，拒绝原假设。Z 得分计算公式如 2-2 所示：

$$Z(I)=\frac{I-E(I)}{\sqrt{V(I)}} \quad (2\text{-}2)$$

其中 $E(I)$ 代表期望值，$V(I)$ 指样本方差。

（3）局部空间自相关

全局空间自相关主要研究整个空间序列的空间聚集情况，在分析传染病空间分布状态时，可以通过局部自相关分析空间异质性[19]，进一步了解某些特定区域的空间集聚情况。局部空间自相关可以用局部 Moran's 指数表示，公式表达如 2-3 所示。

$$I_i = \frac{Z_i}{S_i}\sum_{j\neq i}^{n} w_{ij}Z_j \quad (2\text{-}3)$$

其中，$Z_i = y_i - \bar{y}$，$Z_j = y_j - \bar{y}$，$S^2 = \frac{1}{n}\sum(y_i - \bar{y})$，$w_{ij}$ 是空间权重矩阵，n 是研究区域中所有地区总数，I_i 指第 i 个地区的局部 Moran's 指数。此外，Moran's I 散点图和 LISA 集聚图将不同空间关系更加直观地展示出来，Moran's I 散点图显示所有研究区域所处的四种不同的集聚状态。相同属性值集聚表示存在正的空间自相关（第一、第三象限），反之表示存在负的空间自相关（第二、第四象限）。LISA 集聚图可以显示是否在统计意义上显著，Moran's I 散点图和 LISA 集聚图结合进行分析，可以更加清楚地得出疾病发病率的集聚情况。

2.2 发病率预测方法及评价标准

传染病一般都有较明显的趋势效应和季节效应，研究传染病的关键一步就是利用已有的历史数据外延预测未来变化趋势，考虑传染病数据特性，时间序列分析方法被广泛用于传染病预测。时间序列分析中的 ARIMA 模型只需要考虑序列自身的变化，不涉及其他影响因素。在实际应用中很多学者通过 ARIMA 模型预测发病率，利用实证分析验证了 ARIMA 模型在传染病预测中的适用性和有效性，合理的预测结果对未来的疾病防控提供了依据。

2.2.1 SARIMA 模型

对于单变量时间序列，一般可以通过自回归模型、移动平均模型等进行建模，这些模型在时间序列分析中的应用较为广泛。$ARIMA(p,d,q)$ 模型结构和 $ARMA(p,d)$ 模型很相似，$ARMA(p,d)$ 模型的结构如下所示：

$$x_t = \phi_0 + \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \cdots - \theta_q\varepsilon_{t-q} \quad (2\text{-}4)$$

其中 $\phi_p \neq 0, \theta_q \neq 0; E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t\varepsilon_s) = 0, s \neq t; E(x_t\varepsilon_t) = 0, \forall s < t$，$ARMA$ 模型用延迟算子可以简记为 $\Phi(B)x_t = \Theta(B)\varepsilon_t$。$ARIMA(p,d,q)$ 模型结构为 $\Phi(B)\nabla^d x_t = \Theta(B)\varepsilon_t$，$\nabla^d = (1-B)^d$。当 $d = 0$ 时，$ARIMA(p,d,q)$ 模型就是 $ARMA(p,q)$ 模型。当 $d = 1, p = q = 0$ 时，此时的 $ARIMA(0,1,0)$ 模型也叫做随机游走模型或醉

汉模型[31]。

时间分析的目的就是用历史序列数据预测未知趋势，只有时间序列的特征统计量具有代表性和可延续性，才能保证时间序列的特征统计量不会随时间而变化，即该序列是平稳时间序列。如果原始序列是非平稳序列，可以通过有效地差分或对数化处理使序列平稳。在具体建模过程中，首先要检验时间序列是否为平稳序列和白噪声序列，这些检验是建模的基础，时间序列经过预处理满足建模的条件时，才能进行下一步的模型构建。

（1）平稳性检验

时间序列平稳性的检验，一种方法是图检验，根据已有数据的时序图，直接观察其波动情况判断序列是否存在趋势性。此外，根据自相关图可进一步检验序列平稳性。另一种方法是单位根检验，该方法应用较为广泛，最初始的是 DF 检验，该检验适用范围较小，所以大部分情况下都使用 DF 的修正检验 ADF 检验。对于 $AR(p)$ 过程 $x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \varepsilon_t$，其特征方程为 $\lambda^p - \phi_1 \lambda^{p-1} - \cdots \phi_p = 0$，若 $\left|\lambda^i\right| < 1, i = 1, 2, \cdots p$，特征根在单位圆中，$x_t$ 是平稳序列。

（2）白噪声检验

在实际建模过程中需要进行两次白噪声检验，第一次是对原始序列进行白噪声检验，若序列是白噪声序列，该序列值之间不存在相关关系，后续的建模就无需进行。第二次是模型拟合过程，检验拟合模型的残差是否为白噪声序列，以此说明序列中的线性关系提取充分，模型精度较好。检验统计量如下所示。

$$LB = n(n+2)\sum_{k=1}^{m}(\frac{\hat{\rho}_k^2}{n-k}) \qquad (2\text{-}5)$$

其中 n 是序列观测期数，m 代表延迟期数。通过建立原假设和备择假设检验模型显著性，考虑平稳序列的短期相关性，一般延迟期数选择前几期进行检验即可。

除了上述检验过程，整个建模中的关键一步就是模型识别，确定模型结构。了解 ARIMA 模型的具体形式可以发现，在构建模型的过程中，需要确定相应模型 $ARIMA(d, p, q)$ 中的 p 、d 、q 分别是多少。模型识别过程和具体阶数的确定，可参考表 2.1 中的规律。

表 2.1 ARMA 模型 ACF 和 PACF 与 lag 关系

	自相关系数	偏自相关系数
$ARMA(p,0)$	拖尾	p 阶截尾
$ARMA(0,q)$	q 阶截尾	拖尾
$ARMA(p,q)$	拖尾	拖尾

论文针对乙肝发病率数据本身特性，考虑该序列中蕴含的趋势效应，最终构建了乘积季节模型 $ARIMA(p,d,q)(P,D,Q)_S$。通过一系列检验最终确定模型后，用极大似然法进行参数估计。除了以上两种检验外，在建模中还需要进行参数显著性检验，通过 t 统计量判断模型的参数是否显著。最终在顺利通过所有检验的模型中，选出最优模型。整个建模步骤如下所示：

第一步：根据时序图、自相关图及偏自相关图，观察其是否存在趋势效应，判断序列是否平稳，通过差分等预处理使其变为平稳序列。

第二步：模型定阶，根据自相关系数和偏自相关系数变化情况，找到合适阶数。

第三步：参数显著性检验和残差白噪声检验。检验未通过重新选择模型。

第四步：最优模型的选择，依据为 AIC、BIC 准则，进一步预测序列变化趋势。

2.2.2 BP 神经网络模型

神经网络由基本的神经元构成，这些神经元之间相互联系[2]。神经网络的构造受到了生物的神经网络运作机制的影响，是基于大脑进行信息处理的基础上模拟出来的。早在 1943 年[4] 就有人研究神经网络并提出了“M-P 神经元模型”，该模型主要通过多个输入神经元传递信号，这些神经元被连接到不同的权值上，输入值和神经元的阈值进行对比，经过函数处理后输出。激活函数增加了神经网络模型的非线性，让原本的输入从线性组合变成了非线性形式。常用的激活函数有 sigmoid 函数和 tanh 函数，它们的具体形式如 2-6、2-7 所示。

$$\text{sigmoid 函数：} y=\frac{1}{1+e^{-x}}=\frac{e^{x}}{e^{x}+1} \tag{2-6}$$

$$\text{tanh 函数：} \quad \tanh x = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2-7}$$

在后续的发展中，神经网络方法和技术有了很大突破，对数据的处理越来越智能，其中 BP 神经网络模型被广泛应用并解决了很多实际问题。BP 神经网络比感知机的学习能力更高一个层次，它是一个由误差分布算法训练的多级反馈网络[43]。BP 神经网络包括输入层、隐含层和输出层。在整个训练过程中，首先将训练集从输入层输入，然后到达隐含层，最后传递到输出层进行输出，如果输出误差没有达到我们的要求，此时的神经元就会反向传播，不断调节神经元的阈值和连接权重，直到达到我们所需的条件为止。

论文用 BP 神经网络预测乙肝发病率时，通过前三年月发病率数据预测后一年同期月发病率，设置输入层数为 3，输出层数为 1，将隐含层取值设定在 1-9 之间，依次进行训练，最终根据验证集误差最小确定隐含层数。具体建模步骤如下：

（1）神经网络构建，根据研究的已知数据，判断选取合适的网络结构。

（2）根据已构建网络结构进行训练过程，分析输入值和输出值确定节点数、连接权和阈值，给定学习率和激励函数并不断学习，直到训练后的数据符合要求，否则返回继续训练学习。

（3）网络预测，根据训练好的模型预测后期数据并输出结果。

2.2.3 组合预测模型

单一模型在预测过程中都有各自的优点和缺陷，预测效果可能不尽人意。在此基础上，很多学者提出了组合模型的想法，组合模型主要有两种组合方式，一种是通过给不同种模型各自的权重进行组合，新模型就是不同权重的几种模型将各自预测结果合理结合的方式，这种组合方法要想预测效果好，各模型间要有互补的效应。另一种组合方式是首先用一种模型对原始序列进行预测，预测结果和原始数据存在一定的误差，对于误差采用另一种模型进行处理，最终将两种预测结果相结合，得到最终的预测数据。

论文主要通过 Rstudio 软件实现所有建模过程，ARIMA 建模中使用 forecast 包中的 Arima 函数建立模型；在 BP 神经网络建模中，使用弹性反向传播（RPROP）

算法[6]训练神经网络并进行参数优化，RPROP 是一种快速反向传播学习的直接自适应方法。通过误差函数对权重更新值进行调整。损失函数为 SSE，激活函数选择 logistic 函数，使用阈值为 0.005，学习率为 0.1。

2.2.4 模型评价指标

为了进一步比较各模型的预测效果，论文选取以下几个指标来度量模型的精度。设真实值和预测值分别为 $y=\{y_1,y_2,\cdots y_n\}$，$\hat{y}=\{\hat{y}_1,\hat{y}_2,\cdots \hat{y}_n\}$，$\bar{\hat{y}}$ 和 $\bar{y}$ 分别是 $\hat{y}$ 和 y 的平均值，指标的计算公式如下所示：

绝对误差：$$AE=|y_i-\hat{y}_i| \tag{2-8}$$

相对误差：$$RE=\frac{|y_i-\hat{y}_i|}{y_i} \tag{2-9}$$

均方误差：$$MSE=\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2 \tag{2-10}$$

平均绝对百分比误差：$$MAPE=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right| \tag{2-11}$$

Theil 不等系数：$$U=\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i^{\,2}}+\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^{\,2}}} \tag{2-12}$$

偏倚比例=$$\frac{(\bar{\hat{y}}-\bar{y})^2}{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2/n} \tag{2-13}$$

一般常用的模型预测效果评价标准有 MAPE，MSE。事实上，对模型的评价应该从多个层面展开，首先是对真实值接近程度的评价，其次是对误差均值、误差方差和相对拟合程度的评价。以上指标不仅反映了预测值和真值的平均偏离程度，还从预测值均值和真实值均值的偏离程度考虑，综合来说，上述各指标达标就是一个相对优良的模型。

3 全国乙肝发病流行病学特征

3.1 数据来源

乙肝发病率等相关数据来源于“公共卫生科学数据中心”平台，考虑论文从全国尺度展开分析，所以主要收集整理我国 31 个省市区(不含香港、澳门特别行政区和台湾地区，下同)的乙肝疫情数据。这些数据包含了多种维度下的发病情况。时间涵盖了 2004-2017 年共 14 年，不同省市区的人口数据来源于各自的统计年鉴，此外，获取全国省域基础地理数据，便于后续从空间角度分析乙肝流行特征。

3.2 全国乙肝疫情概况

2004-2017 年我国乙肝发病总数共有 14545089 例，在这个时间段内，乙肝每年的发病数均达到 90 万例以上，其中 2009 年发病数最多。乙肝死亡数呈先增加后逐渐减小的趋势，平均每年死亡数为 655 例，在 2006 年死亡数达到最大值 995 例，2013 到 2014 年死亡数减少速度较快，从 2005-2017 年间乙肝死亡数量减少了一半左右，具体趋势见表 3-1。

表 3-1 2004-2017 我国乙肝发病概况

时间	发病数	死亡数	发病率(1/10 万)	死亡率(1/10 万)	人口(万人)
2004	916426	783	70.50	0.0602	129988
2005	982297	908	75.57	0.0699	130756
2006	1109130	995	84.82	0.0761	131448
2007	1169946	854	89.00	0.065	132129
2008	1169569	831	88.52	0.0629	132802
2009	1179607	792	88.82	0.0596	133450
2010	1060582	689	79.46	0.0516	134091
2011	1093335	637	81.54	0.0475	134735
2012	1087086	582	80.68	0.0432	135404
2013	962974	550	71.12	0.0406	136072
2014	935702	360	69.05	0.0266	136782
2015	934215	352	68.57	0.0258	137462
2016	942268	405	68.74	0.0295	138271
2017	1001952	425	72.61	0.0308	139008

2004-2017 年我国乙肝年均发病率 77.47/10 万，发病率的波动范围在 68.57/10 万到 89.00/10 万之间。这期间乙肝发病率变化趋势可以大致分为三个阶段：2004-2007 年发病率上升阶段，乙肝发病率在 2007 年达到最高值 89.00/10 万；2008-2015 年乙肝发病率呈下降趋势，2015 年下降到最低值 68.57/10 万；2016-2017 年发病率略有上升。与此同时，乙肝的死亡率在 2006 年达到最高值，随后开始逐渐降低，死亡率的最高值和最低值分别为 0.076/10 万和 0.026/10 万。

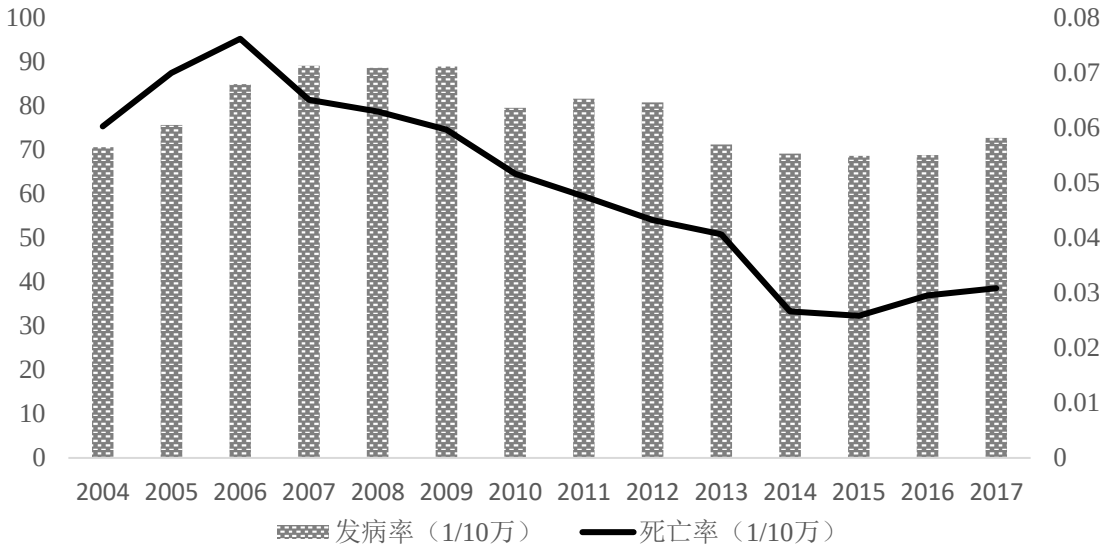


图 3-1 2004-2017 我国乙肝发病概况

3.3 乙肝发病率分布特征

3.3.1 时间分布

乙肝的发病数总体变化趋势是先增加后减少，从 2004 年到 2009 年，发病数由 916426 例增加到 1179607 例，有明显上升趋势，这段时间发病数的增长与传染病信息直报系统的使用有很大关系，实时录入传染病数据，省去了逐级上报的繁琐过程，使传染病数据更完善。从 2010 年开始发病数逐年下降，到 2015 年发病数下降为 934215 例，2016 年乙肝发病数相比上一年出现缓慢上升趋势，在 2017 年总发病数上升为 1001952 例。2004 年至 2017 年全国乙肝发病数的时间分布情况如表 3-2、3-3 所示。

表 3-2 2004-2010 年乙肝月发病数（例）

	2004	2005	2006	2007	2008	2009	2010
1 月	73230	83497	84031	104534	104564	94761	103554
2 月	76637	61918	87581	78025	86401	106561	74025
3 月	86133	89046	103085	109163	111830	112962	103075
4 月	83752	87922	97780	103040	106040	106486	92059
5 月	83237	88127	96789	101560	104349	102377	93190
6 月	78682	83352	95683	99382	98419	103245	87924
7 月	82969	85170	98812	105735	105932	107312	90290
8 月	83967	89043	100779	107469	100842	102891	89898
9 月	70055	79384	87898	90939	88814	88778	79254
10 月	73482	84734	93447	95980	95936	89881	80779
11 月	70738	84362	89284	95056	87915	80652	85579
12 月	53544	65742	73961	79063	78527	83701	80955
总数	916426	982297	1109130	1169946	1169569	1179607	1060582

表 3-3 2011-2017 年乙肝月发病数（例）

	2011	2012	2013	2014	2015	2016	2017
1 月	91764	93334	98292	85776	95359	89176	84430
2 月	81262	107641	72609	76738	66728	75494	89034
3 月	102270	107985	92396	86757	90670	89505	92756
4 月	92251	95234	83468	80656	80342	79298	83305
5 月	95522	99328	84136	80723	80583	83379	87915
6 月	90532	88519	76814	78188	77390	76909	85074
7 月	93702	89987	83092	80601	78918	77194	83303
8 月	96947	88292	81800	79344	77425	80581	85770
9 月	84355	77957	72631	71158	70964	70468	76990
10 月	88993	83492	75216	74393	73045	73572	76826
11 月	84355	80658	72242	70228	71724	73959	79043
12 月	85238	74659	70278	71140	71067	73959	77506
总数	1093335	1087086	962974	935702	934215	942268	1001952

总的来说，2004-2017 年 14 年间，我国平均每年乙肝发病数为 1038935 例。分析每年 12 个月份发病数相加不等于法定报告中的全年发病总数，这可能存在发病数重复计入的问题，所以总数直接根据法定报告中的总数为准。

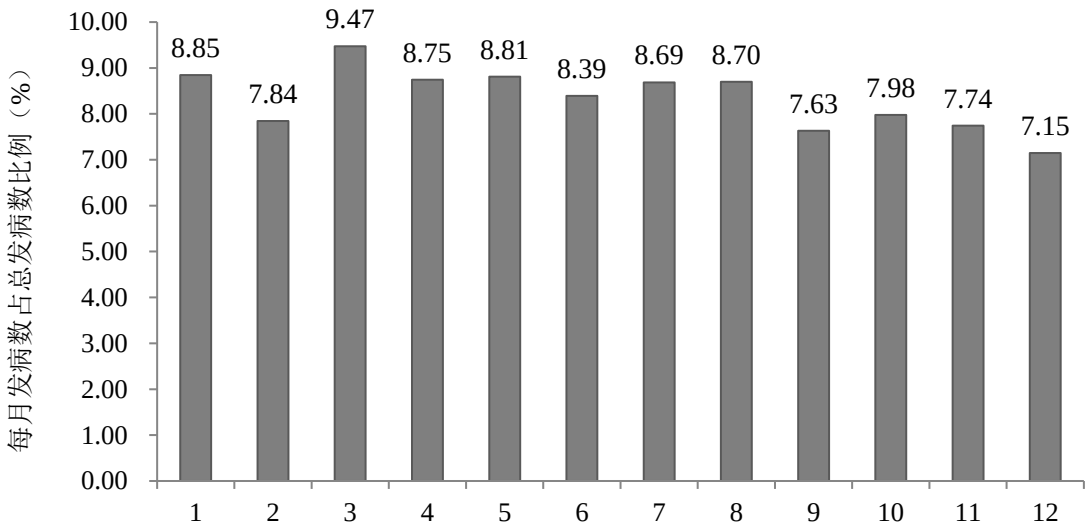


图 3-2 2004-2017 年全国乙肝总发病数分布情况

在 2004 年至 2017 年间， 全国乙肝发病数在每个月份均有分布，2009 年发病数最多，达到 1179607 例。每年 3 月份都是乙肝发病数的高峰期，占总发病数的 9.47%，而 12 月份累计发病数最少，占总发病数的 7.15%，如图 3-2 所示。

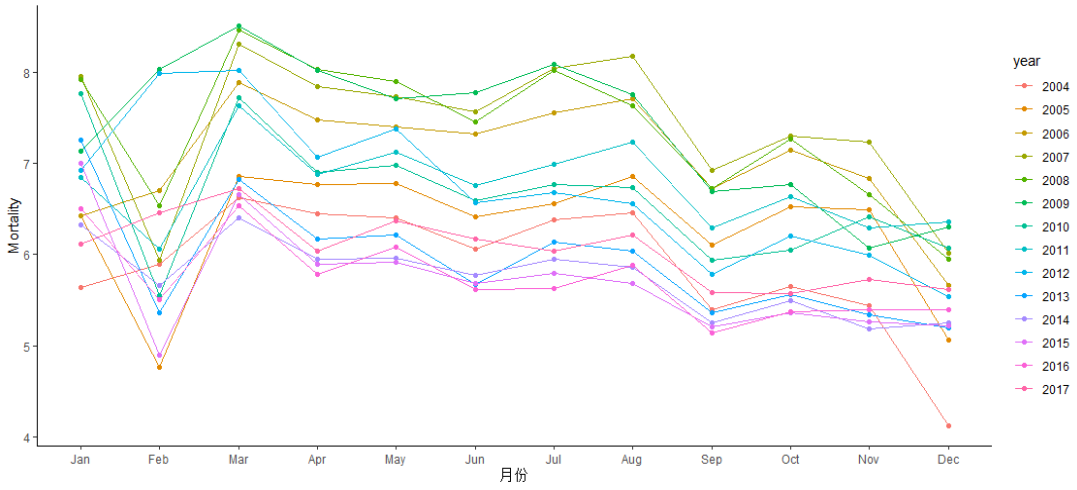


图 3-3 2004-2017 年全国乙肝月发病率分布情况(1/10 万)

2004-2017 年间，我国每年 12 个月份的乙肝发病率变化趋势基本一致，月发病率在 4/10 万-8.5/10 万之间波动，全年月发病率高峰主要出现在 3 月和 8 月，其中 3 月份乙肝发病率达到最高。相比之下，12 月份乙肝发病率最低。

3.3.2 地区分布

论文以省市区为研究单元对全国乙肝发病情况进行分析，根据 31 个省市区 2004-2017 年乙肝发病数据得出，各省市区在不同年度的发病率区别较大，不同地区的年发病率在 7.79/10 万-251.05/10 万之间，地区发病率存在很大的差异。

表 3-4 不同地区乙肝发病数（例）

	2004	2005	2006	2007	2008	2009	2010
全国	916426	982298	1109130	1169946	1169569	1179607	1060582
北京	5579	6999	9593	6671	4564	3821	3172
天津	3428	4866	5899	6167	2829	2421	2164
河北	44109	46280	59282	54985	55485	57309	50062
山西	18109	34192	34826	41172	42132	42747	40903
内蒙古	23177	28091	30246	33698	32645	31976	26474
辽宁	16559	25819	26539	29784	29903	27839	26323
吉林	13989	14436	14819	14889	16761	17820	18192
黑龙江	31291	31551	24209	22740	20672	20240	16390
上海	3548	6385	6208	5857	6264	5704	4654
江苏	21875	18867	18559	17034	15924	14282	12141
浙江	44023	39781	41532	36894	34301	33193	33838
安徽	31437	29515	29976	33233	33224	32256	31548
福建	35637	32961	37353	45151	49100	52644	49173
江西	34778	31871	32065	32102	34249	35784	31817
山东	39563	39733	41835	39799	35027	33569	28274
河南	94243	120984	152043	167146	164408	158728	137445
湖北	55123	53614	59386	69796	73695	79745	69198
湖南	15451	18904	26206	33748	38098	40684	38317
广东	48177	74603	100607	110488	120679	130852	126699
广西	32732	29637	34317	34507	37337	41035	41418
海南	8395	7808	6723	6187	5946	6264	6688
重庆	31311	32381	31343	28988	28093	21038	17317
四川	85560	84643	89742	79944	63345	63091	56327
贵州	18787	18508	24445	30983	32931	35947	28983
云南	13409	18307	23112	21430	19648	20785	21131
西藏	292	308	315	445	466	449	490
陕西	58151	40382	32977	30530	30131	32232	27405
甘肃	49844	49069	59256	64973	65700	62296	53411
青海	9821	11117	11958	14656	17986	22909	14909
宁夏	9303	8928	9162	8735	8522	6794	5435
新疆	18725	21758	34597	47214	49504	45153	40284

续表 3-4 不同地区乙肝发病数（例）

	2011	2012	2013	2014	2015	2016	2017
全国	1093335	1087086	962974	935702	934215	942268	1001952
北京	3116	2612	1944	1648	1683	1696	1900
天津	2075	2039	1713	1702	1898	1854	2148
河北	52550	56785	56248	53177	54593	57835	58207
山西	43812	53117	49608	48518	45433	42355	40009
内蒙古	26892	28170	24661	22566	22028	20672	21351
辽宁	26752	26275	20747	20656	19703	19371	19991
吉林	18012	17115	12996	12710	11245	8435	7454
黑龙江	16549	15451	13675	11439	10230	9336	10819
上海	4662	12556	9843	8551	8426	10054	10711
江苏	13068	12089	10878	11427	14286	14140	12277
浙江	30454	18129	13962	13734	12980	13869	15429
安徽	33379	36659	34752	33930	37275	42729	51432
福建	46524	50058	47247	46577	49254	44008	41296
江西	33828	35387	35093	35167	38715	40795	42421
山东	31026	37923	43615	47439	47722	54732	63258
河南	149852	131568	75256	64429	60032	61107	66772
湖北	67259	63260	59920	58178	61902	63549	66150
湖南	43193	50447	53147	56312	54372	55445	60739
广东	137084	142241	139767	142723	144304	151255	164608
广西	44099	44952	42062	41122	41421	43676	49221
海南	7346	8215	10889	11594	11531	10495	14865
重庆	17723	18612	19190	18476	19753	20373	21266
四川	55314	51591	42627	38686	38573	38688	39488
贵州	21611	20551	20190	19947	20595	17594	20497
云南	22442	25490	24051	25600	21331	13769	14502
西藏	872	981	1239	1733	2500	1927	2187
陕西	26558	28193	25083	25886	23170	20975	17665
甘肃	53098	30722	11862	10511	10047	9897	11042
青海	13118	13456	13249	8509	8536	9557	10632
宁夏	5925	6716	4915	5028	2912	2589	2815
新疆	45142	45726	42545	37727	37765	39491	40800

2004-2017 年全国 31 个省市区的乙肝发病数分布情况如表 3-4 所示，通过观察总体数据可得，广东省发病数最多，共有 1734087 例，占总发病数的 11.92%，而发病数最少的地区是西藏，共有 14204 例，占总体发病数的 0.10%。在研究的 31 个省市区中，2004-2017 年之间发病率最高的地区是新疆，其年均发病率是 178.89/10 万，发病率最低的地区是北京，其年均发病率为 20.82/10 万，从

2004-2017 年甘肃乙肝发病率下降幅度最大，由 190.69/10 万下降到 42.31/10 万。分析各地区乙肝发病率变化情况可知，部分省市区发病率和全国总体发病率发展趋势相同，发病率呈下降趋势，这些省市区分别是北京、吉林、黑龙江、江苏、浙江、重庆、四川、云南、陕西、天津、内蒙古、甘肃、河南、青海和宁夏；而其他省市区的发病率变化趋势与全国总体发病率变化趋势相反，乙肝发病率有所上升。

平均发病率超过全国平均发病率的省市区为：山西、内蒙古、福建、江西、河南、湖北、广东、广西、海南、重庆、陕西、甘肃、青海、宁夏和新疆。其中新疆和青海两地区年均发病率超过 150/10 万，属于发病率最高的地区，江苏、北京和天津年均发病率小于 22.7/10 万，年均发病率最低，如图 3-4 所示。

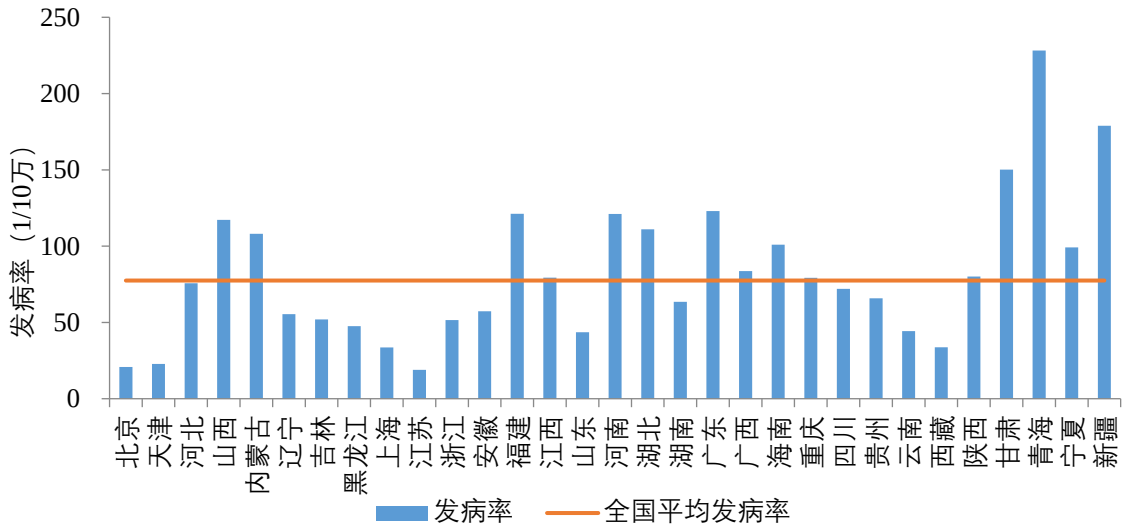


图 3-4 各地区年平均发病率（1/10 万）

根据 2004-2017 年乙肝各地区年度发病率，选取发病数变化幅度较大的部分年份数据展开分析，具体选取 2004、2006、2013、2015、2016 和 2017 年发病率数据。在 ArcGIS 软件中制作出乙肝年度发病率专题地图，更加直观地展示各省市区乙肝发病情况，地图从深绿到深红色，随着颜色深浅的变化，发病率越来越高，具体发病率分布如图 3-5 至 3-10 所示。

从图中可以看出，2004 年乙肝发病率高于 103.10/10 万的区域主要是西部的部分省份，这些省市区分别是甘肃、青海、宁夏和陕西；2006 年发病率高于

103.1/10 万的地区增多，5 个省市区(甘肃、青海、宁夏、新疆和河南)发病率高于 126.76/10 万，发病率均有所上升，部分中部内陆和东部沿海地区的省份发病率升高，如广东、山西。2006 年到 2013 年，乙肝在全国的发病率明显降低，2013 年发病率高于 137.38/10 万的地区有新疆和青海，2015 年中部内陆地区发病率较高，其中包括湖北、湖南、江西等地区，新疆、青海地区仍然是高发病率地区，从 2016-2017 年，发病率有所回升，发病率高于 108.6/10 万的地区有两个，分别是新疆和青海。乙肝发病率下降后没有回升的地区分别是陕西、甘肃、宁夏，四川。2017 年这四个地区发病率均低于 57.65/10 万。吉林、江苏、上海和云南四个地区在 2004-2017 年间发病率均稳定在 65/10 万以下，而湖南、江西和广东地区的发病率逐年升高。总体来说，高发病率范围在 2013 年大幅缩减，但新疆、青海两地区的乙肝发病率一直居高不下。此外，内蒙古、山西、湖北、广西、广东、福建地区发病率也依然较高。

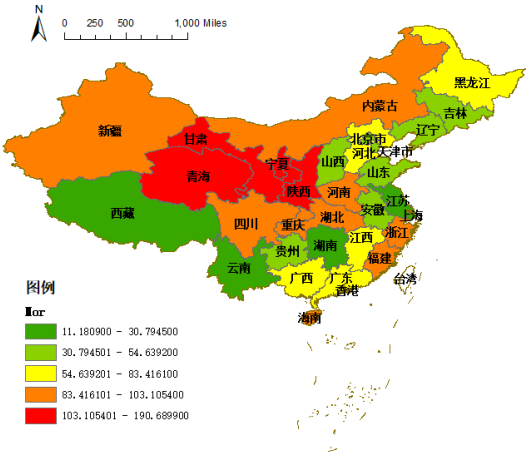


图 3-5 2004 年全国乙肝年度发病率分布

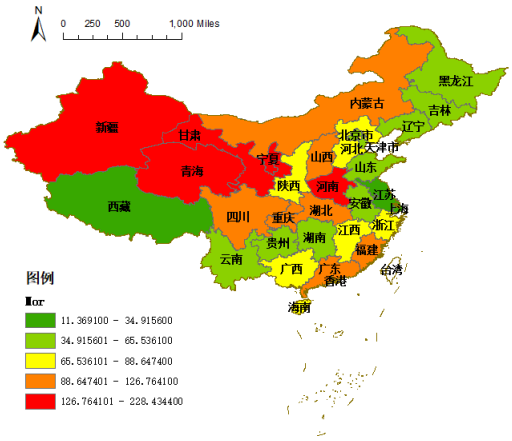


图 3-6 2006 年全国乙肝年度发病率分布

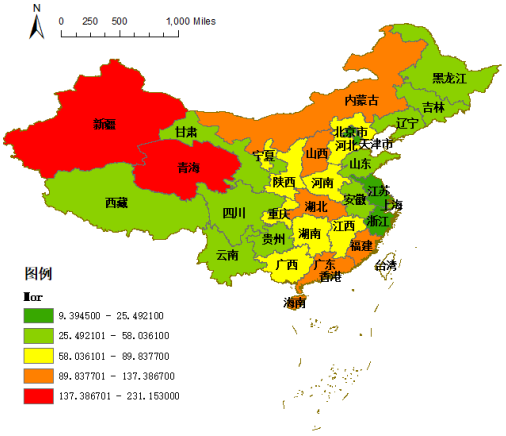


图 3-7 2013 年全国乙肝发病率分布

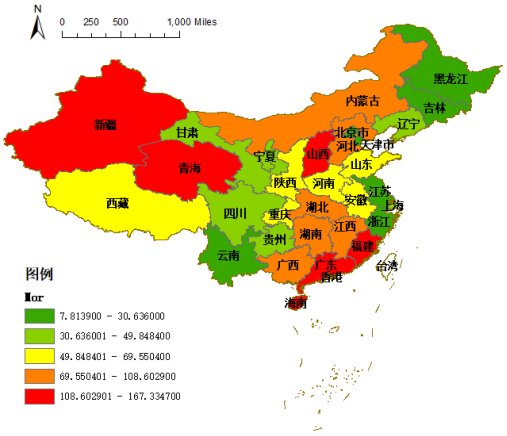


图 3-8 2015 年全国乙肝发病率分布

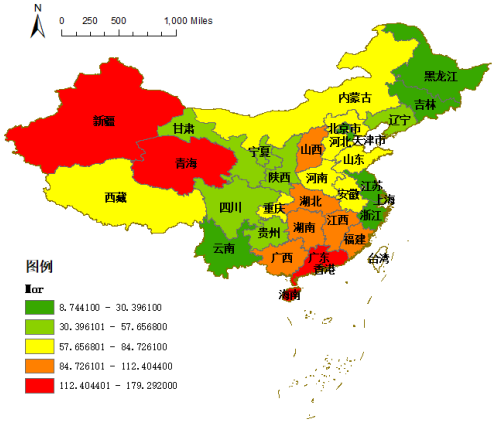


图 3-9 2016 年全国乙肝发病率分布

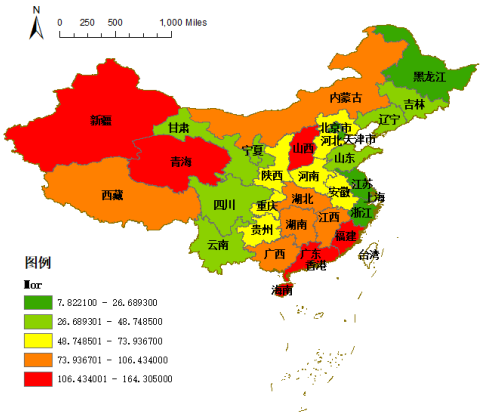


图 3-10 2017 年全国乙肝发病率分布

将我国 31 个省市区划分为东部、中部、西部三个地区，分析乙肝发病率和发病数分布情况及变化规律，具体划分结果如表 3-5 所示。

表 3-5 东、中、西地区划分

东部	中部	西部
北京、天津、山东、浙江、上海、江苏、广东、海南、辽宁、河北、福建、吉林、黑龙江	湖南、内蒙古、湖北、江西、山西、河南、陕西、安徽、	甘肃、云南、四川、重庆、广西、新疆、宁夏、贵州、西藏、青海

表 3-6 东、中、西地区乙肝发病数（例）

年份	东部	中部	西部
2004	316173	330469	269784
2005	350089	357553	274656
2006	393158	397725	318247
2007	396646	441425	331875
2008	397455	448582	323532
2009	405958	454152	319497
2010	377770	403107	279705
2011	389218	424773	279344
2012	401488	426801	258797
2013	383524	357520	221930
2014	383377	344986	207339
2015	387855	342927	203433
2016	397080	347627	197561
2017	422963	366539	212450

2004-2012 年，中部地区乙肝发病数高于东部地区，2013-2017 年，东部地区乙肝发病数最多，中部和西部地区乙肝发病数变化趋势保持一致，都呈现先增加后减少的趋势，而东部地区发病数增加，无明显下降趋势。2004-2017 年中部地区发病总数最多，共有乙肝发病数 5444186 例。2009 年东部地区和中部地区乙肝发病数最多，分别有 405958 例和 454152 例。

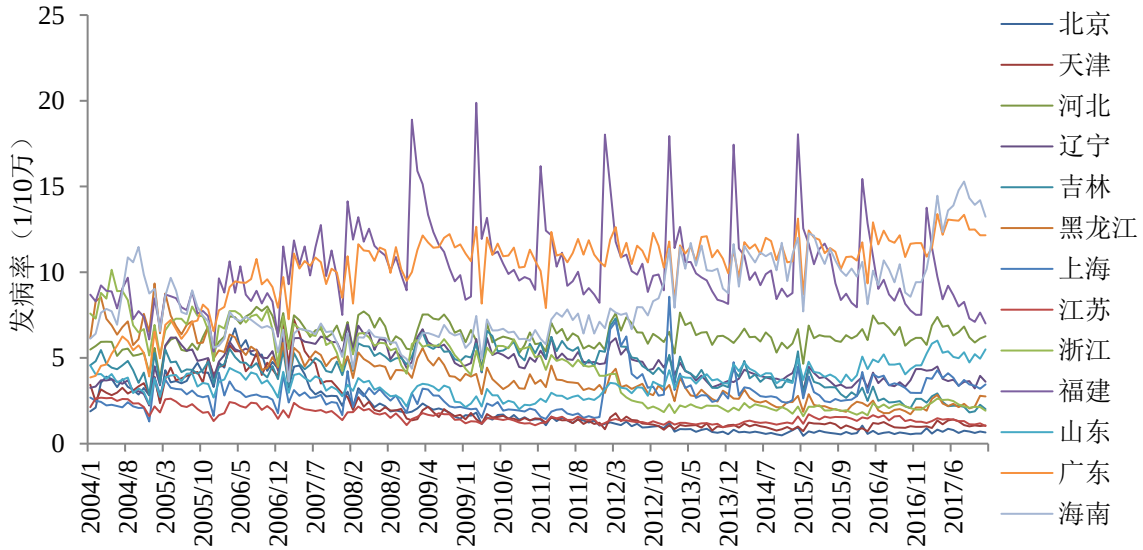


图 3-11 东部地区发病率变化趋势

从2004-2017年间，东部大部分省份的乙肝发病率均保持在10/10万以下，且发病率与全国总体发病率变化趋势一致，逐年下降；而广东、海南和福建乙肝发病率呈明显上升趋势，其中福建变化幅度最大，发病率最高达到19.88/10万。

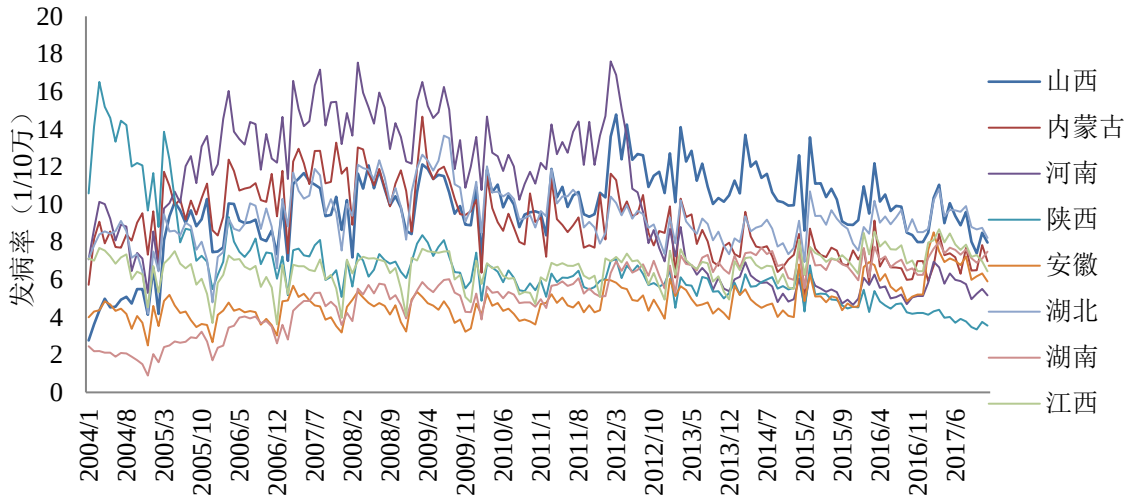


图 3-12　中部地区发病率变化趋势

随着时间变化，西部地区大部分省份乙肝的月发病率逐渐趋向于 7/10 万左右；根据发病率变化趋势可分为两种：其中陕西、河南、内蒙古乙肝发病率呈下降趋势，其他 5 个省份的月发病率有上升趋势。

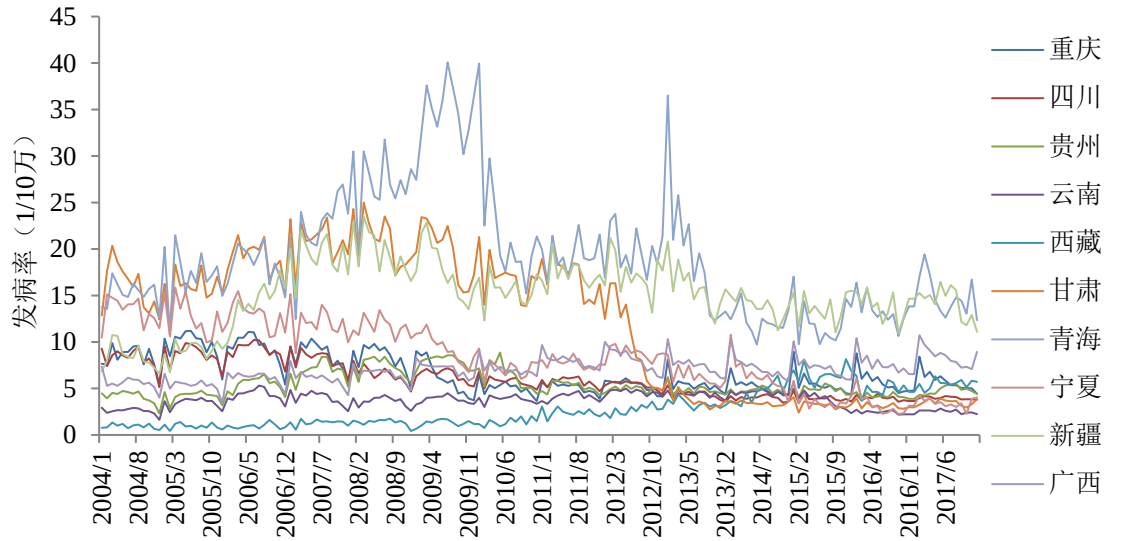


图 3-13　西部地区发病率变化趋势

甘肃乙肝的月发病率变动幅度最大，发病率明显下降，其次青海下降幅度较大，西藏、云南、宁夏、四川、重庆、贵州乙肝的月发病率始终在 10/10 万左右，从 2004 年到 2017 年，月发病率均逐渐趋于 5/10 万左右。

综上所述，西部地区和中部地区的乙肝发病率波动越来越小，各省市区发病率变化幅度减小，而东部地区各省市区的乙肝发病率波动越来越来，发病率趋势线越来越分散。

3.3.3 人群分布

（1）性别分布

男性乙肝发病数和发病率始终高于女性，2004-2017 年男性乙肝总发病数为 9312490 例，女性乙肝总发病数为 5232599 例，分别占总发病数的 64.02%和 35.98%。在这段时间范围内，男性和女性乙肝平均发病率分别为 96.58/10 万和 57.29/10 万，2004 年至 2017 年发病数中男女性别比分别为 2.02、1.97、1.88、1.85、1.81、1.79、1.76、1.68、1.67、1.67、1.68、1.74、1.73、1.75。

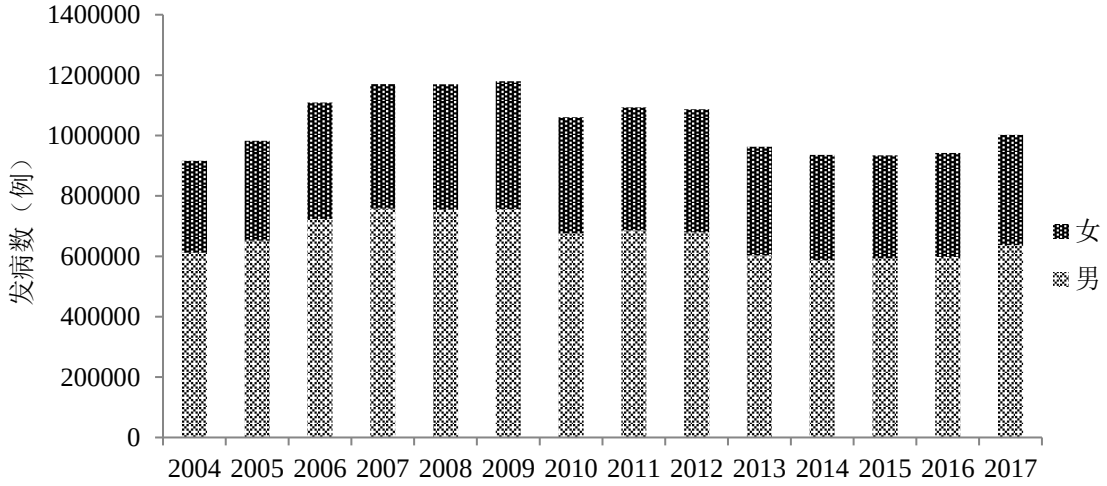


图 3-14 2004-2017 年不同性别的乙肝发病数

总体来看，我国乙肝患病人群中男性占了很大比例，14 年间的男性乙肝发病率始终高于女性乙肝发病率。2007 年男性发病率达到最高值 111.64/10 万，而女性发病率在 2009 年达到最高值，为 65.49/10 万，随着时间的变化，男性和女

性的乙肝发病率都逐渐下降，且发病数中的男女性别比也有所下降。具体分布情况如表 3-7 所示。

表 3-7 2004-2017 年不同性别乙肝发病分布

	男		女	
	发病数（例）	发病率 (1/10 万)	发病数（例）	发病率 (1/10 万)
2004	611525	91.59	304901	48.52
2005	651534	96.99	330763	52.34
2006	723374	107.08	385756	60.70
2007	757906	111.64	412040	64.48
2008	753805	110.52	415764	64.70
2009	756398	110.42	423209	65.49
2010	675781	98.37	384801	59.13
2011	685544	99.49	407791	62.25
2012	679325	98.12	407761	61.93
2013	602913	86.67	360061	54.41
2014	587062	83.98	348640	52.41
2015	592930	84.41	341285	51.03
2016	597432	84.60	344836	51.28
2017	636961	89.74	364991	53.94

（2）年龄分布

不同年龄段的乙肝分布情况都不相同，论文将年龄段分为 0-9 岁，10-19 岁，20-29，以此类推，最后将大于等于 80 岁的人群分为一组。在 2004 年到 2017 年所有乙肝发病数中，80 岁以上年龄组发病数最少，一共有 33684 例，只占 14 年总发病人数的 0.23%。其次是 0-9 岁人群，发病数占所有发病数的 1.18%。20-29 岁人群乙肝发病数共有 3351199 例，发病数最多，占总发病人数的 23.18%；其中 40-49 岁、50-59 岁、60-69 岁、70-79 岁发病数占比分别是 21.3%、19.52%、14.49%、8.90%。

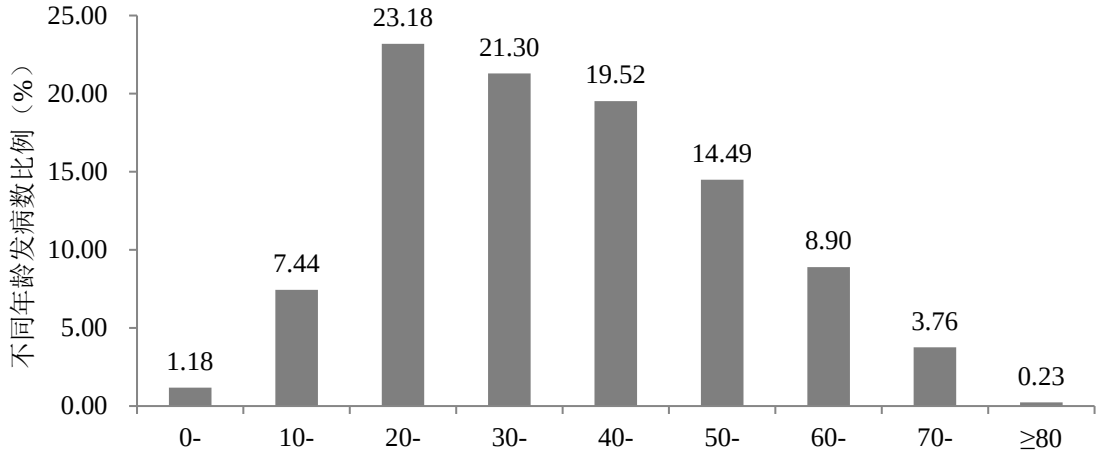


图 3-15　2004-2017 年发病数在不同年龄段占比

分析 2004-2017 年各年龄段乙肝发病趋势可以发现，不同年龄段乙肝发病数发生了明显变化。年龄段在 0-39 岁之间的乙肝发病数呈下降趋势，其中 0-9 岁人群发病数从 25190 例减少到 5316 例，10-19 岁发病数下降幅度最大， 2004 年有乙肝发病数 150015 例，2017 年减少到 19380 例；与此同时，40 岁以上年龄段的发病数均呈现上升趋势，40-49 岁间的乙肝发病数从 137453 例增加为 219598 例，50-59 岁的发病数由 92279 增加到 186043、60-69 岁、70-79 岁、80 岁以上的发病数也分别增加了 92310 例、36002 例和 3692 例。

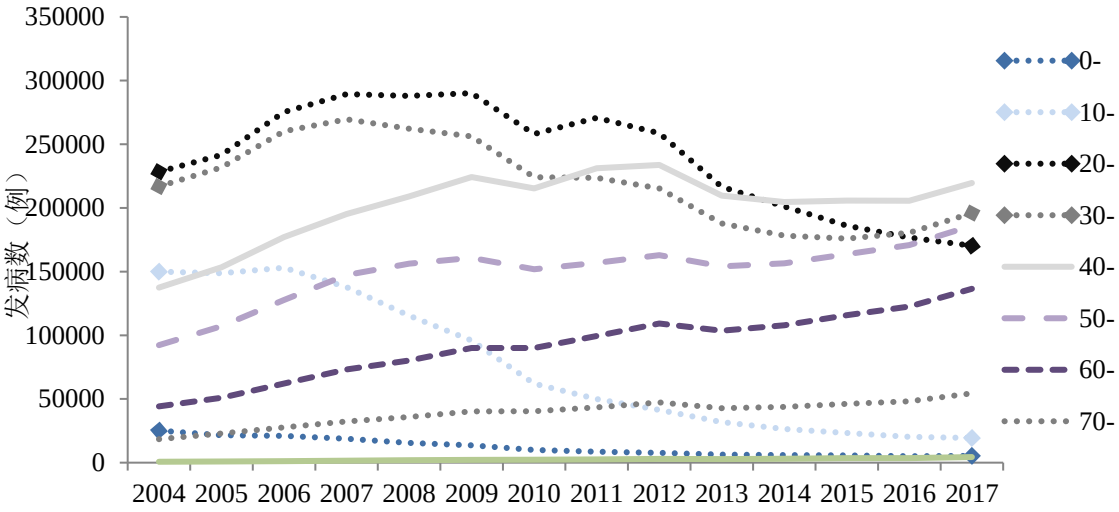


图 3-16　2004-2017 年全国乙肝分年龄发病数

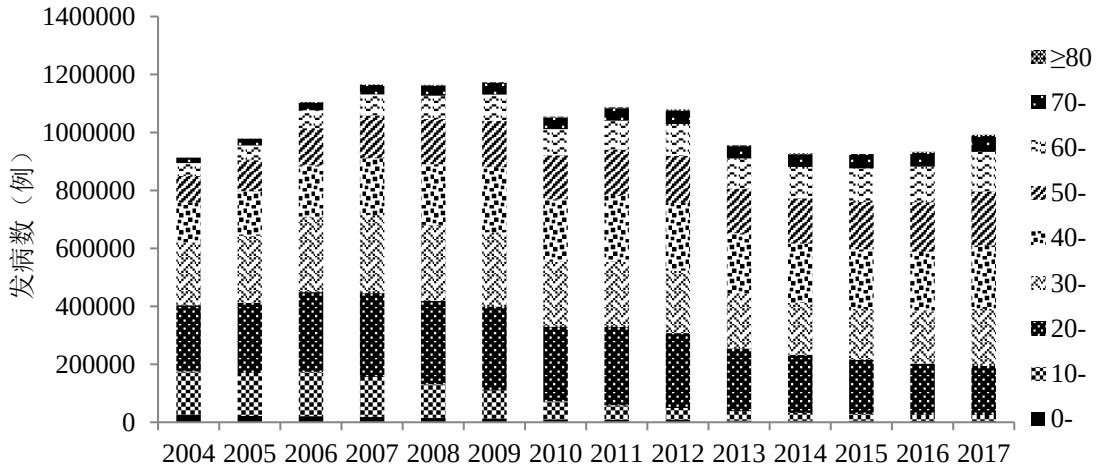


图 3-17 分年龄组发病数分布

总的来说，我国乙肝发病数主要分布在 20-59 岁人群之间，即发病数主要集中在青壮年时期。0-19 岁人群病例数较少，其中 10-19 岁发病数下降 130635 例，该年龄段发病数下降幅度大，这与我们国家的乙肝免疫预防工作紧密相关，充分说明我国新生儿接种乙肝疫苗有效地预防了低年龄人群感染乙肝病毒。与此同时，高年龄人群的发病数却逐年增长。2004-2017 年全国乙肝分年龄发病数情况如图 3-15，图 3-16、3-17 所示。

（3）职业分布

2004-2017 年全国所有乙肝发病数中，职业为农民的发病总数为 7156424 例，占比 49.20%，将近总发病人数的一半。此外，从事家政、家务及待业人群的发病数占总病例数的 10.46%，工人发病总数也较多，在所有职业中发病数排第三，占比 8.61%。其他职业的病例数占比均在 6%以下。不同职业乙肝发病数表明农民和工人等职业的乙肝发病数较多，这也侧面反映了劳动强度、经济水平和医疗设施等因素可能和乙肝分布密切相关，后续应该更加有针对性的对乙肝进行预防和治疗，从这些高发病职业人群出发，做好乙肝易感人群的防疫宣传活动，提高易感人群和高危人群对乙肝病毒危害性的意识，防患于未然。2004-2017 年不同职业乙肝发病数如表 3-8、图 3-18 所示。

表 3-8 2004-2017 我国不同职业乙肝发病数

职业	发病数（例）	构成比（%）
幼托儿童	25870	0.18
散居儿童	79237	0.54
学生	862176	5.93
教师	206527	1.42
保育员	2547	0.02
餐饮食品人员	45405	0.31
公共场所服务员	12231	0.08
商务人员	413417	2.84
医务人员	55591	0.38
工人	1252739	8.61
民工	316179	2.17
农民	7156424	49.20
牧民	67865	0.47
渔(船)民	13842	0.10
海员及长途驾驶员	11913	0.08
公务人员及职员	566047	3.89
离退休人员	605807	4.17
家政、家务及待业	1520997	10.46
不详	672891	4.63
其它	657384	4.52
合计	14545089	100.00

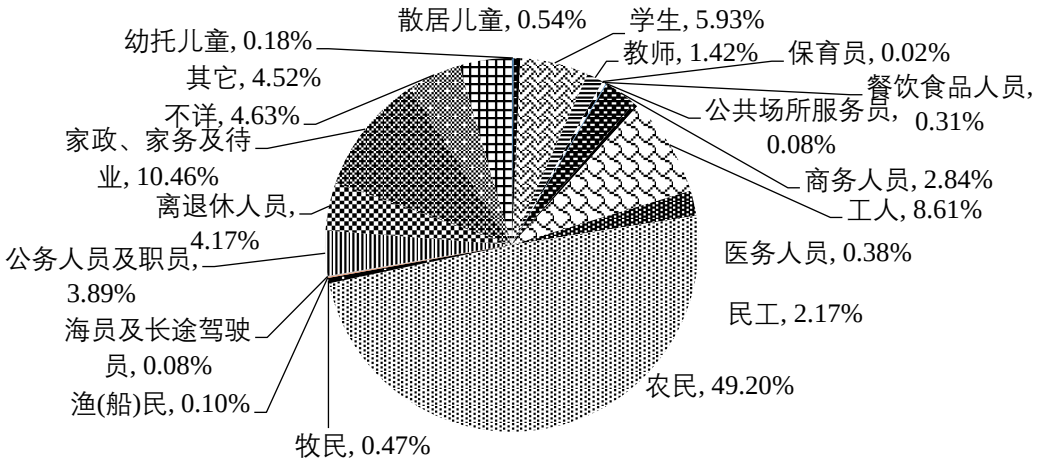


图 3-18 全国 2004-2017 年乙肝发病数在不同职业分布情况

综合以上分析可以得出，我国乙肝的发病具有鲜明的时间特征、地区特征和人群特征。不同月份发病数有差异，3 月份是乙肝发病的高峰期，12 月份发病数最少；地区发病数和变化趋势存在差异，新疆、青海乙肝发病率居高不下，广东

乙肝发病数持续上升；乙肝发病数中，男性占比高于女性；年龄段在 20-59 岁之间的乙肝发病数占总发病数的 78.49%；农民、工人发病数占总发病数一半以上。

3.4 乙肝发病率时空特征分析

结合 ArcGIS 的地图创建功能，建立全国 2004-2017 年的乙肝年均发病率地图，可以更加直观地显示乙肝在全国的时空分布。

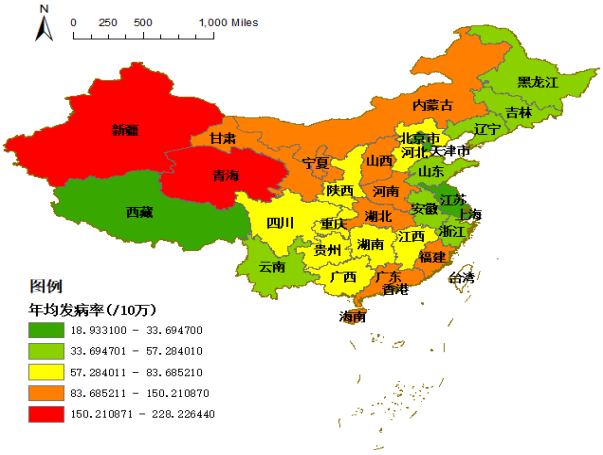


图 3-19　2004-2017 年 14 年间全国各地区乙肝年均发病率专题地图

根据中国 31 个省市区（不含港、澳、台）的空间分布特征，如图 3-19 可看出，各省发病率在空间分布上存在明显的高集聚区和低集聚区。高发病率省份主要位于我国的西北地区，主要包括青海和新疆，这两个地区年均发病率都在 150/10 万以上。西北地区出现这种集聚现象的原因可以从两方面考虑：首先，新疆、青海等高发病率地区均在地理位置上相邻，高发病率地区可能对邻近区域产生影响，其次，西部地区经济发展，卫生条件和医疗设施水平较低，这也可能导致乙肝发病率高于其他地区。发病率的空间分布情况，为研究空间自相关性提供了很好的思路。考虑到全国各省市区乙肝发病率高低有明显集聚性，所以从整体角度把握，分析区域间的相关性和异质性具有现实意义。

3.4.1 空间集聚特征

论文选用 Moran’s I 指数分析我国乙肝发病率的全局空间自相关性，具体操

作使用 GeoDa 软件实现。首先通过各省市区的邻接关系建立空间权重矩阵，当两相邻省市区存在公共边界时，空间权重为 1，反之空间权重为 0。在实际创建空间权重矩阵时需结合我国的行政区域划分，由于海南省没有邻接省份，在研究过程中设定其与广东省邻接；其次根据空间权重矩阵计算全局 Moran's I，得出各省市区乙肝发病率的全局空间聚集效应，具体结果如表 3-9 所示。

表 3-9 2004-2017 年中国省市区发病率全局 Moran's I

年份	Moran's I	标准差	Z 值	P 值
2004	0.369	0.1202	3.3291	0.003
2005	0.343	0.1178	3.1688	0.005
2006	0.315	0.1167	2.9586	0.007
2007	0.327	0.1148	3.0753	0.005
2008	0.312	0.1105	3.1123	0.005
2009	0.245	0.1035	2.6752	0.009
2010	0.284	0.1116	2.8306	0.008
2011	0.338	0.1143	3.2293	0.003
2012	0.270	0.1135	2.6903	0.007
2013	0.214	0.1073	2.3618	0.01
2014	0.306	0.1172	2.9333	0.005
2015	0.337	0.1235	3.0026	0.007
2016	0.297	0.118	2.725	0.007
2017	0.345	0.1226	3.0878	0.007

根据表 3-9 可以看出，从 2004 年到 2017 年，我国乙肝发病率全局 Moran's I 都大于 0，Z 得分为正且 P 值小于 0.05，通过显著性检验，说明各省市区发病率并不是表现为完全的随机性，我国 2004-2017 年乙肝发病率存在显著的空间正相关性，空间联系特征具体表现为：高值(低值)与高值(低值)趋于空间集聚。

3.4.2 局部空间依赖性及异质性

全局空间自相关反映整体范围内的集聚情况，不能检验局部地区是否存在空间相关性和空间异质性，为了进一步研究各省市区发病率的局部特征，论文选取 2004 年、2008 年、2012 年和 2017 年各地区乙肝发病率数据，通过莫兰散点图和 LISA 集聚图，将各区域的空间联系形式可视化展示出来，莫兰散点图可以进一步显示区域单元和其邻接地区间的聚集情况，LISA 图可以显示集聚的显著度，

两者相结合可使分析的空间依赖性和异质性更准确。具体结果见图 3-20。

分析莫兰散点图可得，四种空间集聚形式都存在，其中大部分省市区位于散点图的一三象限，少数分布在二四象限，这说明发病率在我国省市区间既存在空间相关性，又存在空间异质性。

根据 2004 年莫兰散点图可知，一共有 24 个点分布在第一、第三象限中，每个点代表一个地区，一三象限的省市区数占所有地区数的 77.4%。其中第一象限包含的地区分别是新疆、四川、陕西、宁夏、内蒙古、湖北、河南、甘肃、福建和重庆，说明 2004 年高发病率地区主要集中在西北地区和邻近区域。而云南、辽宁、湖北、北京、天津、黑龙江、贵州、安徽、江苏、广东、河北、吉林、湖南、广西处于第三象限，说明低发病率地区主要集中在东部。2008 年，高-高集聚区包含新疆、甘肃、青海、宁夏、山西、内蒙古，低-低集聚区有江苏、上海、天津、云南、浙江、贵州、广西、山东、辽宁、江西、吉林、安徽、河北、北京。高-高值区域个数减少，山西由原来的低-高聚集模式变为高-高聚集模式。2012 年处于高-高值区域的地区有新疆、山西、宁夏、内蒙古、河南、海南、甘肃、青海、广东，新增河南、海南、广东三个省份，说明南部沿海地区也出现发病率高-高聚集的空间分布模式。相比 2004 年，低-高型集聚区域明显增多，这些地区乙肝发病率较低，被周围高发病率地区所包围，而低-低型集聚区域包含北京、天津、浙江、云南、山东、辽宁、吉林、贵州、安徽、上海、重庆、江苏和河北。2017 年只有西藏、河南、甘肃处于低-高型集聚区，说明西藏和甘肃这两个低发病率地区被新疆和青海高发病率地区所包围，河南发病率比周围的湖北省发病率低。与此同时，黄河中游地区的山西、河南，内蒙及相邻区域不再处于高-高集聚区，高-高集聚区出现了长江中游的湖北、湖南、江西三个地区。

综上所述，我国乙肝发病率在空间上主要呈现出高-高型和低-低型两种状态，一三象限的地区总数没有很大变化，但是这些象限中的省市区不断发生改变。从地理位置可以看出，最开始高-高聚集区主要在我国的西北地区和相邻区域，随着时间的推移，从北向南，逐渐有南部沿海地区、黄河中游地区和长江中游地区的部分省份落入高-高集聚区。与此同时，青海始终属于乙肝的高发病地区，位于莫兰散点图的第一象限，与青海邻近的地区大部分也处于高发病率状态。由此可得，空间上相邻地区的乙肝发病率具有相似的属性，且发病率的空间差异性较

小，我国乙肝发病率一直存在空间相关性和空间异质性。

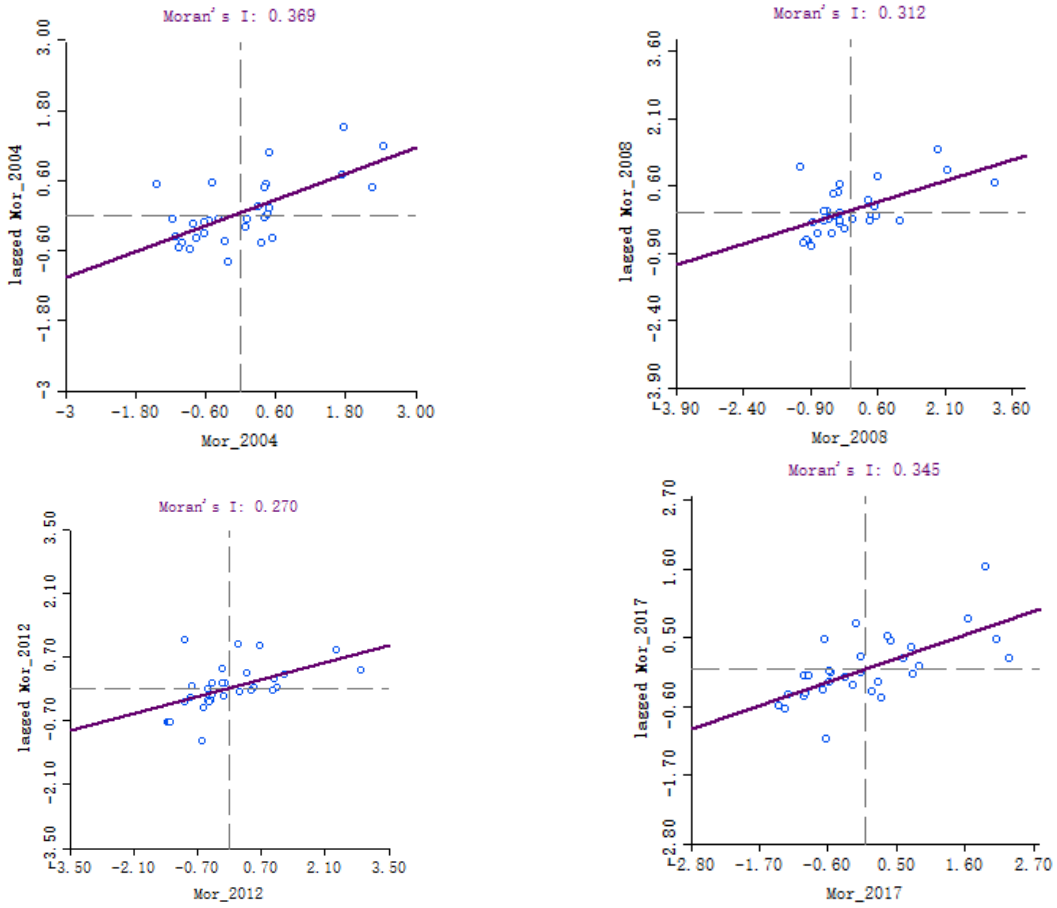
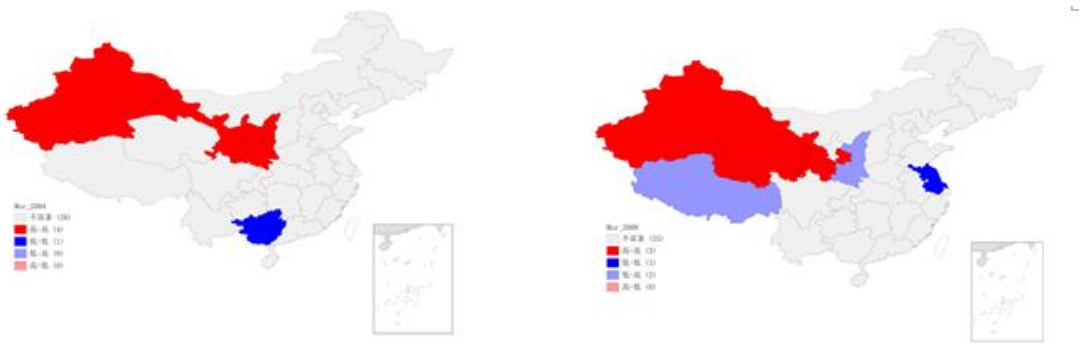


图 3-20　2004、2008、2012 及 2017 年乙肝发病率局部 Moran's 散点图

上述 Moran's I 散点图并没有对各个省市区的局部 Moran's I 指数进行检验，论文通过使用 LISA 图进一步衡量各省乙肝发病率和邻接省域的相关程度，具体结果如图 3-21 所示。



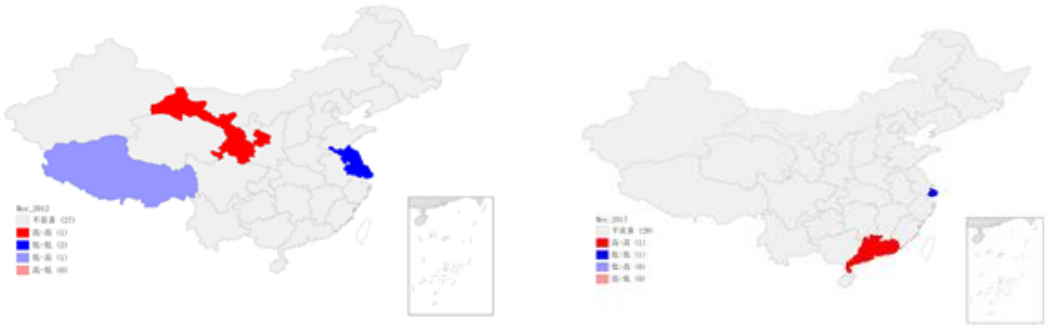


图 3-21 2004 年、2008 年、2012 年和 2017 年中国各省乙肝发病率集聚情况

LISA 图中的红色区域和蓝色区域表示乙肝局部空间自相关分析具有统计学意义，红色是乙肝高发病地区，蓝色区域表示发病率低。2004 年已经具有明显的高-高集聚效果，集聚区主要包含西北地区的新疆、甘肃、宁夏、陕西，表明这几个地区是我国乙肝的热点区域。与此同时，低-低集聚地只有广西。2005 年只出现了高-高集聚地，且聚集区域仍然是新疆、甘肃、宁夏、陕西；2006 年陕西从高-高集聚状态跨越到低-高集聚状态，新增江苏地区位于低-低集聚区；2007-2008 年，除了低-高集聚区新增西藏地区，其他聚集形式所包含的地区均比较稳定；2009 年陕西从低-高集聚变为不显著；2013 年四种集聚形式的地区数发生了明显变化，只出现了低-低集聚和低-高集聚两种空间集聚形式，低-低值分布的地区上海，说明本地乙肝发病率低，周围邻近省市区发病率高。低-高值分布的西藏和甘肃乙肝发病率低于周围邻近省市区的发病率；2014 年-2017 年，四种集聚状态均保持不变，其中高-高集聚区为广东，低-低集聚区为上海。

综上所述，高-高集聚区的省市区数目最多，低-高集聚区的省市区数目次之，低-低集聚区的省市区数目最少，高-高集聚区主要分布在我国西北地区的大部分省份，之后又有部分地区位于高-高集聚区，如南部的广东，低-低集聚区主要出现在东部沿海地区，包括江苏和上海。

4 全国乙肝月发病率预测分析

4.1 单一模型的建立与预测

4.1.1 ARIMA 模型的构建

（1）数据平稳化

根据 2004-2016 年全国乙肝月发病率数据，绘制发病率的时间序列图。通过观察时序图特征，初步判断序列是否存在趋势效应。为了更加清晰地展示乙肝发病率的特征，进一步将时间序列拆分为趋势成分、季节成分和随机成分 3 种不同成分展开分析。根据趋势图可知，乙肝发病率呈先上升后下降的趋势，季节图显示该序列存在季节性趋势。由此可得乙肝发病率序列不满足平稳性要求，具体结果如图 4-1 和 4-2。

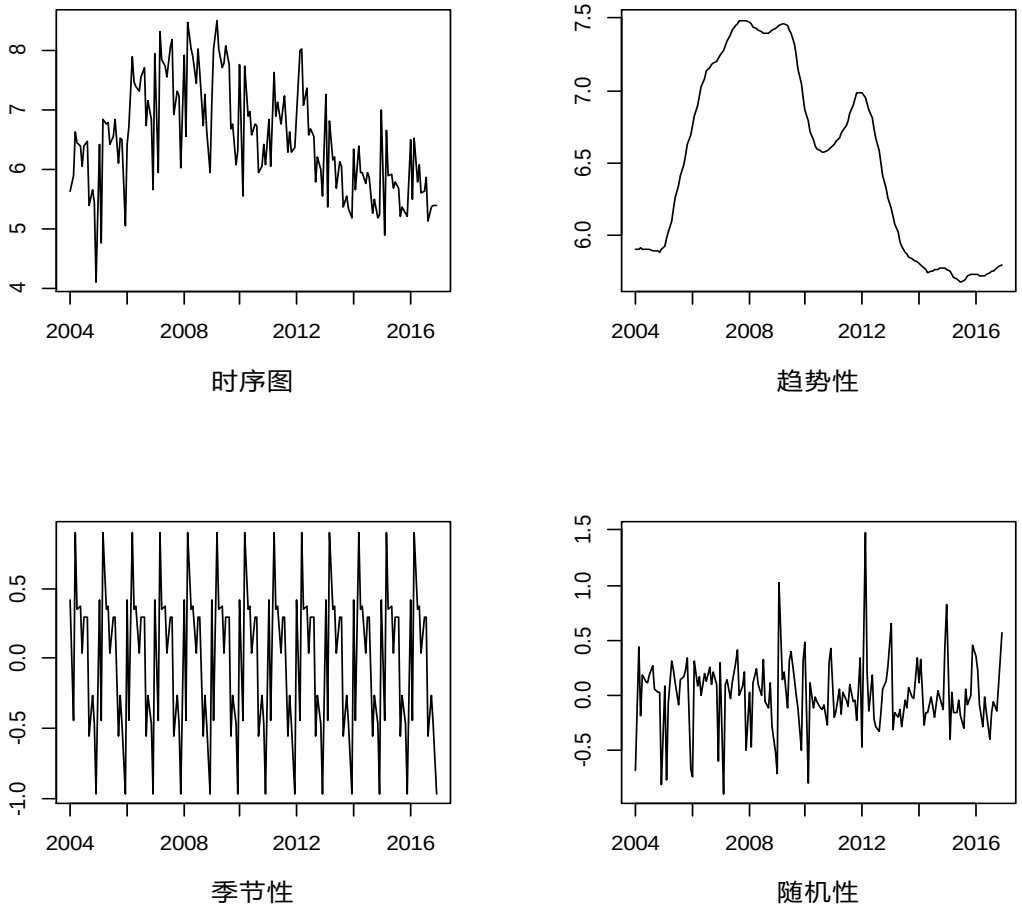


图 4-1　2004 年 1 月-2016 年 12 月我国乙肝发病率的时间分布

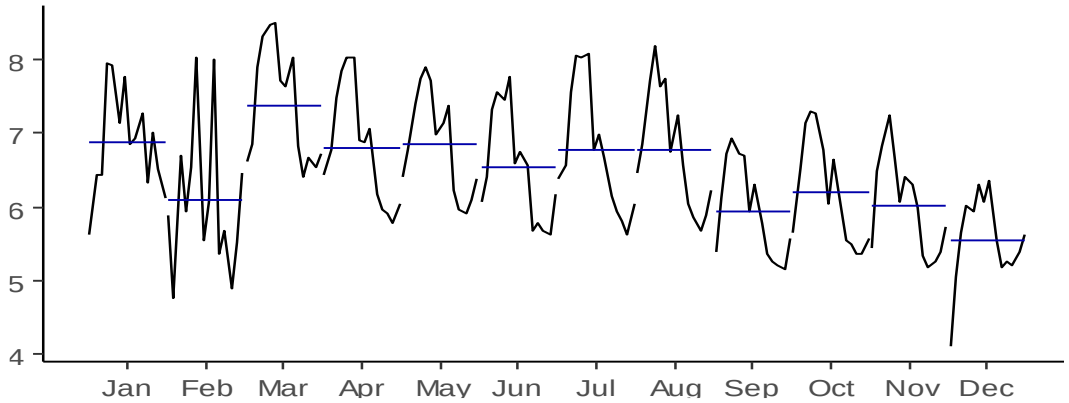


图 4-2 我国乙肝发病率(1/10 万)的月份分布

上述分析得出乙肝发病序列是一个非平稳序列，通过一阶差分和一阶季节性差分消除乙肝发病率序列的不平稳性。差分运算比较充分地提取了原序列中的长期趋势和季节效应，其结果表明序列显示出平稳序列特征。根据 2.3.1 建模步骤，进一步检验序列随机性，白噪声检验结果如表 4-1 所示，延迟不同阶数之后的 p 值均远远小于显著性水平 0.05，该结果表明差分后的时间序列不是白噪声序列。检验前 6 期和 12 期延迟是考虑平稳序列的短期相关性，同时避免期数太长淹没序列的相关性[31]。具体结果如图 4-3 所示。

表 4-1 序列白噪声检验

延迟阶数	χ^2 统计量	p 值
1	45.798	1.31E-11
6	46.897	3.10E-08
12	118.34	2.20E-16

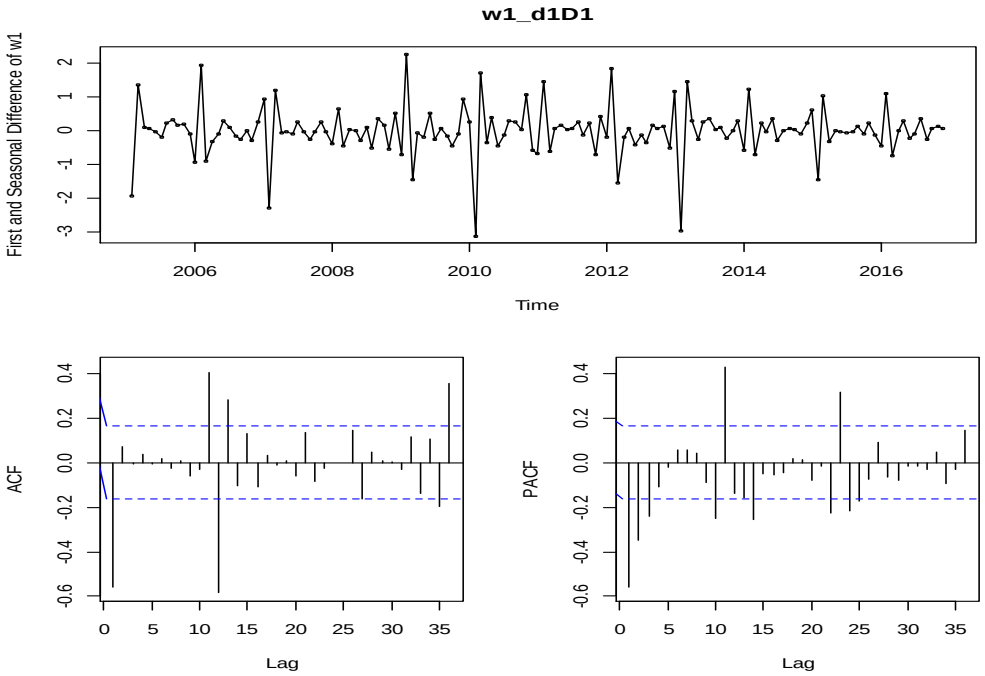


图 4-3　2004 年 1 月-2016 年 12 月我国乙肝发病率 ACF 和 PACF 图（差分后）

（2）模型拟合

乙肝发病率的自相关图和偏自相关图显示，2004-2017 年的乙肝月发病率不是纯随机时间序列，且该序列存在季节性趋势，考虑平稳性检验中的趋势差分阶数和周期为 12 的季节差分阶数，ARIMA 模型的阶数确定为 $d=1, D=1$。根据乙肝发病率序列的自相关图可知，自相关系数在延迟一阶之后，迅速减小，可初步考虑 $q=0$ 或 1；偏自相关系数变化情况表明 p 的阶数为 0 到 3 之间；P 由偏自相关系数在 12 阶的性质决定，自相关系数在 2 阶后迅速衰减到 2 倍标准差置信区间以内，偏自相关系数在一阶后仍拖尾，所以进一步推出 $P=0, Q=2$。

表 4-2　ARIMA 模型 AIC 指标结果

ARIMA 模型	log likelihood	AIC
(1,1,2)(0,1,2)12	-75.32	173.2
(2,1,1)(0,1,2)12	-74.92	162.99
(2,1,0)(0,1,2)12	-75.72	162.93
(2,1,2)(0,1,2)12	-74.88	164.65
(3,1,0)(0,1,2)12	-75.13	164.53

根据以上分析建立所有可能的模型，通过对比模型的 AIC 值选取最优模型。

对 2004-2016 年月发病率数据进行拟合最终得出乙肝发病率数据的 ARIMA 模型，具体模型为 $ARIMA(2,1,0)(0,1,2)_{12}$。

（3）模型检验

首先对模型残差进行白噪声检验，通过 Ljung-Box 统计量，得出不同延迟阶数对应的 p 值均大于显著性水平 0.05，说明 $ARIMA(2,1,0)(0,1,2)_{12}$ 模型的残差序列为白噪声序列。此外，根据 t 统计量检验参数显著性可知，估计所得系数除以标准差的绝对值均大于临界值 1.96。综上所述，论文拟合的 ARIMA 模型显著有效，为下一步乙肝发病率的合理预测奠定了基础。

表 4-3 ARIMA 模型参数估计

	Estimates	Std Error	t
AR	-0.7468	0.0789	9.465
AR	-0.3723	0.0823	4.524
SMA	-0.8981	0.0783	11.470
SMA	0.3867	0.0926	4.176

4.1.2 基于 ARIMA 模型的预测

根据 ARIMA 建模步骤,得到最终模型为 $ARIMA(2,1,0)(0,1,2)_{12}$，利用已构建 ARIMA 模型拟合 2004-2016 年的乙肝发病率，得到图 4-4 的拟合图。

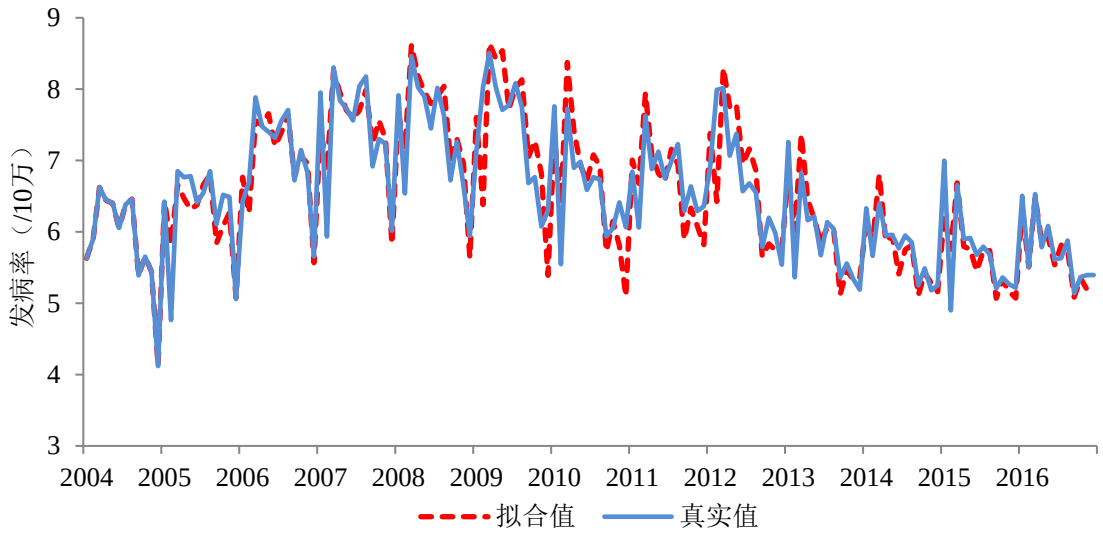


图 4-4 2004-2016 年乙肝发病率拟合效果图

该图显示拟合值和真实值的整体走向基本一致，较好地反映了乙肝发病率的整体变化趋势，进一步用已有模型对 2017 年 1 月-12 月的乙肝发病率进行内插预测，模型预测结果和真实值进行对比，验证模型的可靠性，具体结果如表 4-4 所示：

表 4-4 2017 年乙肝发病率(1/10 万)真实值和预测值对比

时间	真实值	预测值	绝对误差	相对误差
2017/1	6.1188	6.7045	0.5857	9.57%
2017/2	6.4525	5.2551	-1.1974	18.56%
2017/3	6.7222	6.6255	-0.0967	1.44%
2017/4	6.0373	6.1441	0.1068	1.77%
2017/5	6.3714	6.4877	0.1163	1.82%
2017/6	6.1655	5.8590	-0.3065	4.97%
2017/7	6.0372	6.1515	0.1143	1.89%
2017/8	6.2159	6.1591	-0.0568	0.91%
2017/9	5.5796	5.5827	0.0031	0.06%
2017/10	5.5678	5.7492	0.1814	3.26%
2017/11	5.7284	5.5860	-0.1424	2.49%
2017/12	5.6170	5.5164	-0.1006	1.79%

ARIMA 模型内插预测结果显示，2017 年真实值的月平均发病率为 6.05/10 万，预测的月平均发病率为 5.98/10 万。3 月份仍是一年的高发病时期，1 月份和 2 月份预测值和真实值误差较大，其他月份的预测值绝对误差都在 0.3 以下，相对误差控制在 5%以下，平均相对误差为 4.04%，模型整体拟合效果比较好。

4.1.3 BP 神经网络模型构建

（1）数据选取

已知 2004-2017 年发病率原始数据，根据前三年历史月发病率数据预测下一年同期乙肝发病率，最终可得到 2004-2017 年发病率样本数 132 个，输入值可以表示为 X_1、X_2、X_3（如 2004 年 1 月发病率、2005 年 1 月发病率、2006 年 1 月发病率）,输出值为 Y（2007 年 1 月发病率）。选取 2007-2016 年的数据作为训练集拟合模型，2017 年 1-12 月发病率数据作为测试集评估模型。

（2）神经网络模型参数确定

数据集分割完成后，根据研究需要设置网络层数和各层神经元数目。论文选择只有一个隐含层的 BP 神经网络，避免隐含层过多导致模型复杂且训练耗时。考虑论文输入数据特征，将输入层和输出层的节点数量分别设定为 3 和 1。把隐藏层神经元个数设为M，范围是 1-9，依次训练并进行参数优化，激活函数设定为 logistic。

4.1.4 基于 BP 神经网络模型的预测

数据处理及模型设定完成之后，将前三年月发病率数据作为输入值，后一年数据作为输出值，2017 年 1 月之前的数据为训练集，2017 年全年的数据作为测试集。

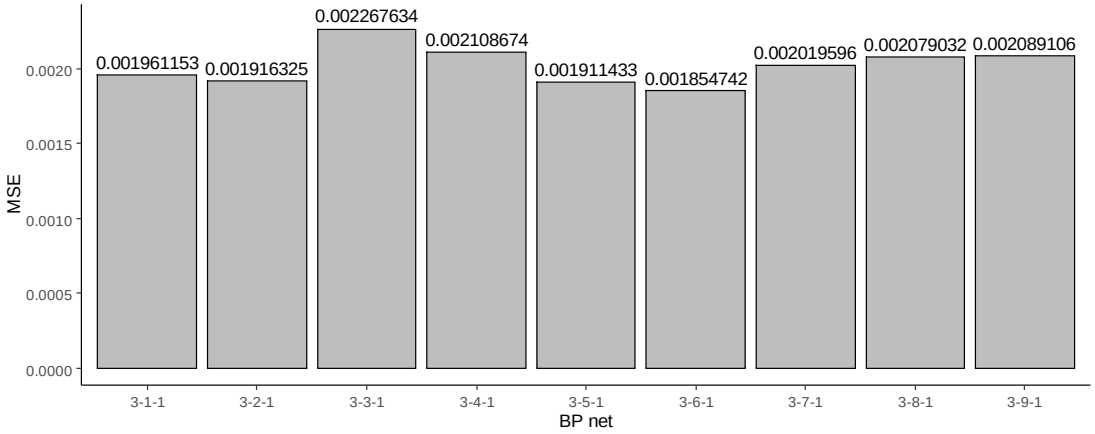


图 4-5　9 种神经网络结构的均方误差

将训练集和验证集输入模型运行，训练过程中不断修正权重值和误差值，最终根据测试集实际值和期望值的均方误差大小，确定最优神经网络。训练结果如图 4-5 所示，随着网络结构变化，测试集的均方误差各不相同，选择均方误差最小值对应的网络结构 3-6-1。整个模型训练过程的网络结构如图 4-6 所示。

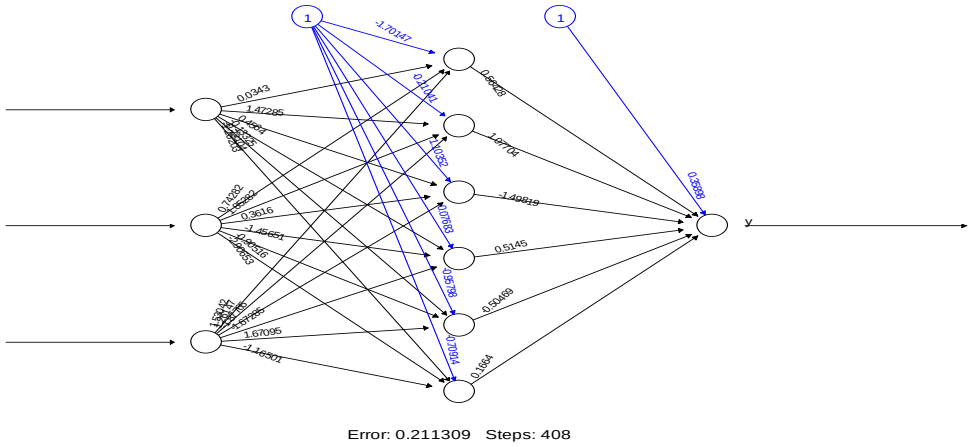


图 4-6 BP 神经网络结构

根据神经网络输出层数据，得到发病率预测值，绘制模型拟合曲线检验拟合效果，具体结果见图 4-7。由拟合效果图可得，蓝色实线代表的真实发病数据逐年下降，红色虚线代表的 BP 神经网络模型拟合值的整体变化趋势与实际发病率一致。

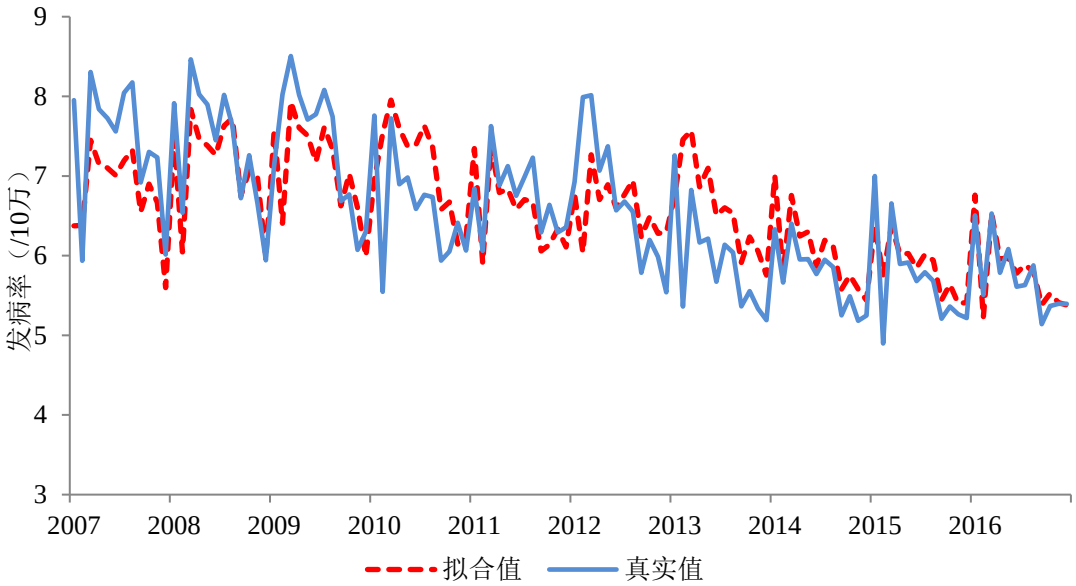


图 4-7 2004-2016 年乙肝发病率拟合效果图

用测试集测试网络性能，如果模型对训练样本以外的数据预测效果好，则该模型外推有效。通过确定的最优神经网络模型，将测试集输入并运行，进一步得

到神经网络模型预测的 2017 年乙肝发病率数据，对比实际发病率和预测数据，具体结果如表 4-5 所示。

表 4-5 2017 年乙肝发病率（1/10 万）真实值和预测值对比

时间	真实值	预测值	绝对误差	相对误差
2017/1	6.1188	6.5742	0.4554	7.44%
2017/2	6.4525	5.7862	-0.6663	10.33%
2017/3	6.7222	6.5200	-0.2022	3.01%
2017/4	6.0373	6.1494	0.1121	1.86%
2017/5	6.3714	6.1984	-0.1730	2.71%
2017/6	6.1655	6.0327	-0.1328	2.15%
2017/7	6.0372	6.0993	0.0621	1.03%
2017/8	6.2159	6.0912	-0.1247	2.01%
2017/9	5.5796	5.7309	0.1513	2.71%
2017/10	5.5678	5.8565	0.2887	5.18%
2017/11	5.7284	5.7729	0.0445	0.78%
2017/12	5.6170	5.7747	0.1577	2.81%

由表 4-5 中的预测值可以得出，在预测的 12 个月份中，除了 1 月份和 2 月份相对误差较大之外，其他预测数据的相对误差都控制在 5%以内，平均相对误差为 3.50%。真实发病率的平均水平为 6.051/10 万，预测值的平均水平为 6.049/10 万。预测值和真实值都很接近，对比单一 ARIMA 模型，BP 神经网络模型预测效果更好。

4.2 组合模型的建立与预测

传染病预测模型种类越来越多，选择合理模型预测对疾病防控有重要指导意义。ARIMA 模型作为单一模型，它在时间变量中蕴含了影响发病的信息，可以充分提取发病序列中的周期性和趋势性，最终获得的模型可以预测发病率的线性趋势，但忽略了序列中的非线性趋势。而 BP 神经网络有强大的非线性映射能力，可以更好的拟合发病序列中的非线性部分。ARIMA-BP 组合模型结合两种模型的优势，其综合考虑了发病率中的季节趋势和非线性趋势，通过串联组合的方式将两种模型组合起来，可以获得更佳的拟合效果。

4.2.1 ARIMA-BP 模型构建流程

上述通过单一 ARIMA 模型和神经网络模型对发病率进行预测拟合，整体拟合效果较好，但仍有部分数据误差较大，预测结果不稳定，所以综合考虑两种单一模型优势对数据信息提取特征。时间序列通常都有趋势效应，在具体的分析过程中，同时考虑序列的线性部分和非线性部分。论文将时间序列和机器学习方法相结合，先通过 ARIMA 模型拟合时间序列的线性部分，将最终预测结果记为$\hat{L}_t$，然后用 BP 神经网络模型预测非线性部分，即 ARIMA 的残差序列，记为$\hat{e}_t$，用串联组合的方式将两种模型组合起来，进一步优化拟合效果，整个过程不涉及权重问题。模型预测结果为线性部分和非线性部分预测结果相加，具体模型构造可参照上述 4.1 中两种单一模型进行，建模流程图如图 4-8 所示。

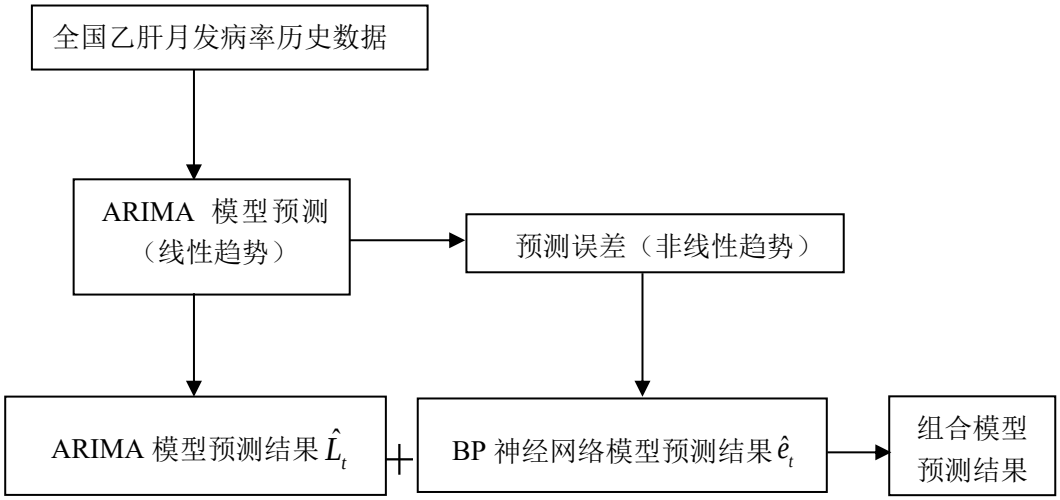


图 4-8 组合模型建模流程图

4.2.2 模型的拟合

首先利用 2004-2016 年全国乙肝月发病率数据构建 ARIMA 模型，提取乙肝发病率中的线性、周期性及季节性信息，具体构建过程和 4.1.1 中的单一 ARIMA 模型相同，通过数据处理和模型识别，得到最优模型 $ARIMA(2,1,0)(0,1,2)_{12}$，模型预测结果反映了乙肝发病率序列中的线性趋势。ARIMA 模型预测乙肝发病率获

得的残差序列如表 4-6 所示。

表 4-6 ARIMA 模型残差序列

时间	1	2	3	4	5	6	7	8	9	10	11	12
2004	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01
2005	-0.02	-1.06	0.19	0.30	0.47	0.03	-0.11	0.05	0.27	0.44	0.22	-0.02
2006	-0.34	0.43	0.33	-0.02	-0.25	0.12	0.16	0.07	-0.05	0.09	-0.13	0.09
2007	0.49	-0.92	0.06	-0.11	0.02	-0.03	0.34	0.19	-0.31	-0.26	-0.07	0.19
2008	0.24	-0.62	-0.15	-0.16	-0.07	-0.35	0.12	-0.41	-0.28	-0.04	-0.30	0.30
2009	-0.47	1.64	-0.14	-0.41	-0.83	0.05	0.06	-0.38	-0.37	-0.51	-0.75	0.91
2010	0.45	-0.85	-0.65	-0.53	0.05	-0.13	-0.31	-0.18	0.21	-0.10	0.60	0.97
2011	-0.16	-0.62	-0.32	-0.19	0.30	0.04	-0.17	0.29	0.40	0.31	0.21	0.54
2012	-0.45	1.56	-0.28	-0.68	-0.42	-0.38	-0.48	-0.31	0.13	0.37	0.24	-0.21
2013	0.45	-0.66	-0.54	-0.29	0.03	-0.18	0.08	0.03	0.24	0.09	0.01	-0.16
2014	0.20	-0.21	-0.40	0.09	0.01	0.36	0.20	0.04	0.14	0.07	-0.10	0.09
2015	0.78	-0.86	-0.03	0.10	0.16	0.23	0.06	-0.05	0.14	0.04	0.10	0.15
2016	0.17	0.00	0.04	-0.07	0.15	0.07	-0.18	0.13	0.06	0.02	0.21	0.27

构建 BP 神经网络模型训练学习上述残差序列，BP 模型的构建过程参考单一 BP 模型，由于样本数较少，仍然构建只有单一隐含层的神经网络模型。针对 2004-2016 年的残差数据，用前三年数据预测第四年同期数据，以此生成训练集，2017 年残差数据作为测试集。论文同样设定输入层节点数为 3，输出层为 1，隐含层设定取值为 1-9，依次通过训练集训练，直到测试集的实际输出值和期望输出值均方误差最小。

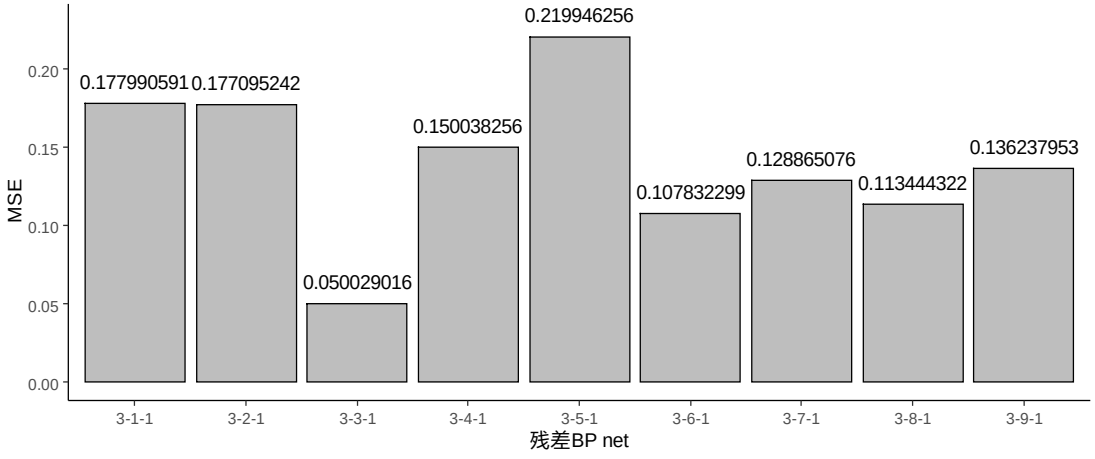


图 4-9 不同神经网络结构的均方误差

通过模型训练和比较，得到隐含层为 1-9 时对应的均方误差,由图 4-9 可知，均方误差最小为 0.05，此时对应的隐含层数为 3。所以最终确定神经网络结构为 3-3-1，利用该神经网络对 ARIMA 模型的残差进行预测，激活函数设定为 logistic。

4.2.3 拟合效果评估

根据建立的 BP 模型预测残差值，残差预测结果和 ARIMA 模型的预测值相加，最终得到组合模型的预测值。绘制发病率的拟合效果图，由图 4-10 可得，红线（拟合值）和蓝线（真实值）的变化趋势一致，两条线重合度高，组合模型拟合效果良好。

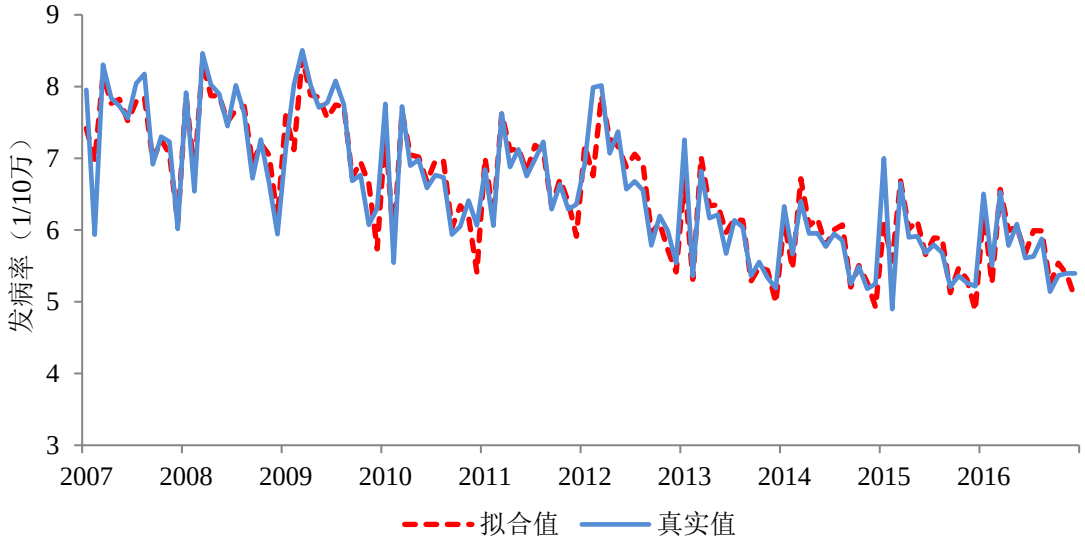


图 4-10 2004-2016 年乙肝发病率拟合效果图

利用构建的 ARIMA-BP 神经网络组合模型预测 2017 年的月发病率，比较预测结果和真实值误差大小，评估组合模型的拟合效果，组合模型得到的 2017 年乙肝发病率分别是 6.53/10 万、5.94/10 万、6.59/10 万、6.13/10 万、6.50/10 万、5.91/10 万、6.15/10 万、6.34/10 万、5.61/10 万、5.80/10 万、5.63/10 万、5.63/10 万。其预测值的平均相对误差为 2.95%，比单一 ARIMA 模型和 BP 神经网络模型平均相对误差小，模型的预测结果更加接近真实发病数据。预测结果见表 4-7。

表 4-7 2017 年乙肝发病率（1/10 万）真实值和预测值对比

时间	真实值	预测值	绝对误差	相对误差
2017/1	6.1188	6.5361	0.4173	6.82%
2017/2	6.4525	5.9449	-0.5076	7.87%
2017/3	6.7222	6.5947	-0.1276	1.86%
2017/4	6.0373	6.1341	0.0968	1.60%
2017/5	6.3714	6.5069	0.1355	2.15%
2017/6	6.1655	5.9104	-0.2551	4.14%
2017/7	6.0372	6.1550	0.1178	1.95%
2017/8	6.2159	6.3460	0.1301	2.09%
2017/9	5.5796	5.6150	0.0354	0.63%
2017/10	5.5678	5.8090	0.2412	4.33%
2017/11	5.7284	5.6311	-0.0973	1.70%
2017/12	5.6170	5.6318	0.0148	0.26%

4.3 预测效果比较及模型选定

论文根据 2004-2017 年乙肝发病率数据特征，选取合理方法预测乙肝发病率，具体使用 ARIMA、BP 神经网络和两种单一模型的组合模型对乙肝发病率数据进行拟合和内插预测。对比不同模型的均方误差、平均绝对误差、平均绝对百分比误差、Theil 不等系数和偏倚比例，根据这些评价指标结果综合评价模型预测效果，选择最优模型预测乙肝发病率的未来趋势。具体评估结果见表 4-8 和图 4-11。

表 4-8 模型预测效果对比

	ARIMA	BP	ARIMA-BP
MSE	0.1654	0.0754	0.0534
MAE	0.2507	0.2142	0.1813
MAPE	4.04%	3.50%	2.95%
Theil 不等系数	0.0337	0.0226	0.0190
偏倚比例	0.0264	0.0001	0.0053

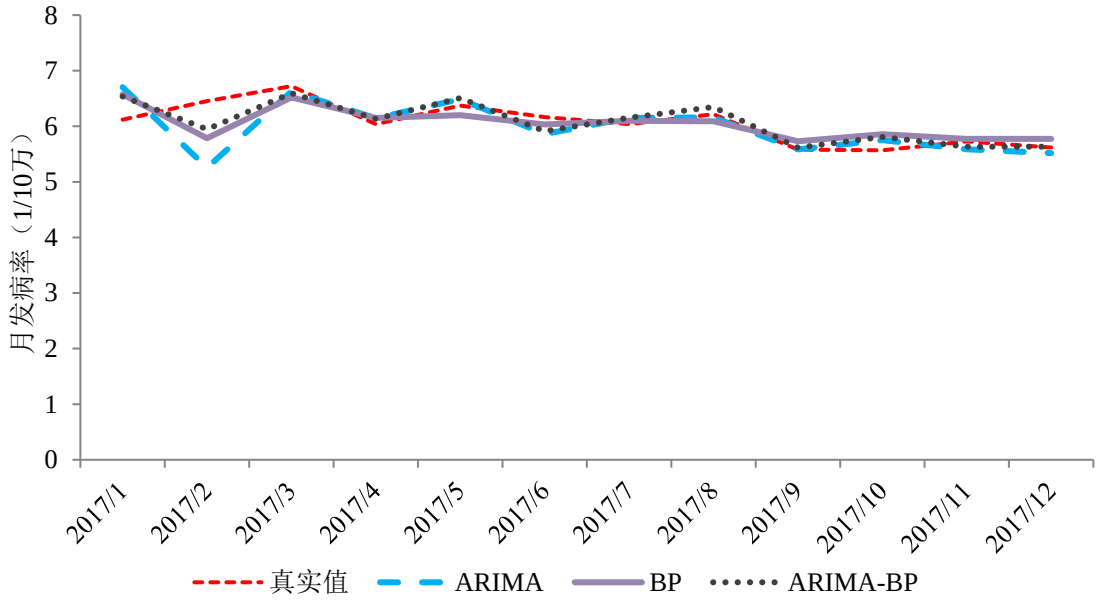


图 4-11　三种模型对乙肝发病率的预测趋势对比

对比真实值和预测值折线图可知，三种模型的预测值和发病率真实值的整体变化趋势都较为贴合，发病率变化幅度小的月份拟合效果更好，对比模型预测效果最差的 1 月份和 2 月份数据可以发现，组合模型 ARIMA-BP 和实际发病数据间的波动更小，相对误差最小，其次是 BP 神经网络模型，ARIMA 模型和真实值误差最大。ARIMA，BP 和 ARIMA-BP 的均方误差分别为 0.1654、0.0754、0.0534，平均绝对误差分别为 0.25、0.21、0.18。Theil 不等系数分别为 0.0337、0.0226、0.0190。相比两种单一模型，用 ARIMA-BP 组合模型的预测值描述真实值时具有更好的精确度。

4.4 预测结果及分析

基于以上分析结果，乙肝发病预测的最优模型是 ARIMA-BP 组合模型，因此用 ARIMA-BP 组合模型对乙肝未来的发病趋势进行预测。第一步，用 $ARIMA(2,1,0)(0,1,2)_{12}$ 模型预测 2018 年的月发病率，得到 1-12 月发病率预测值；第二步，BP 神经网络模型构建参考上面的建模过程，根据 2004-2017 年数据建立的 $ARIMA(2,1,0)(0,1,2)_{12}$ 模型，提取出相应的残差序列。设定含有单一隐含层的

神经网络模型，输入层节点数和输出层节点数分别为 3 和 1，激活函数选择 logistic，根据训练好的 BP 神经网络得到模型非线性部分的预测值，由图 4-8 组合模型建模流程图所示，将组合模型预测值相加，得到最终预测如表 4-9 所示。

表 4-9　2018 年全国乙肝月发病率预测值

时间	ARIMA 预测结果	BP 预测结果	ARIMA-BP 预测结果
2018/01	6.9608	-0.0801	6.8807
2018/02	5.7771	-0.4208	5.3563
2018/03	6.8976	-0.0033	6.8944
2018/04	6.2105	-0.1605	6.0499
2018/05	6.3798	-0.1305	6.2492
2018/06	6.0472	0.3008	6.3480
2018/07	6.1220	-0.1786	5.9435
2018/08	6.1857	-0.1038	6.0819
2018/09	5.5814	-0.1339	5.4475
2018/10	5.7593	-0.1601	5.5992
2018/11	5.7019	-0.0536	5.6483
2018/12	5.6847	-0.1056	5.5791

由表 4-9 可知，2018 年乙肝月发病率区间在 5.45/10 万-6.89/10 万之间，2018 年乙肝的月均发病率为 6.01/10 万。全国乙肝发病率最高的月份是三月份，冬季发病率较低，这与历史实际发病率趋势保持一致。

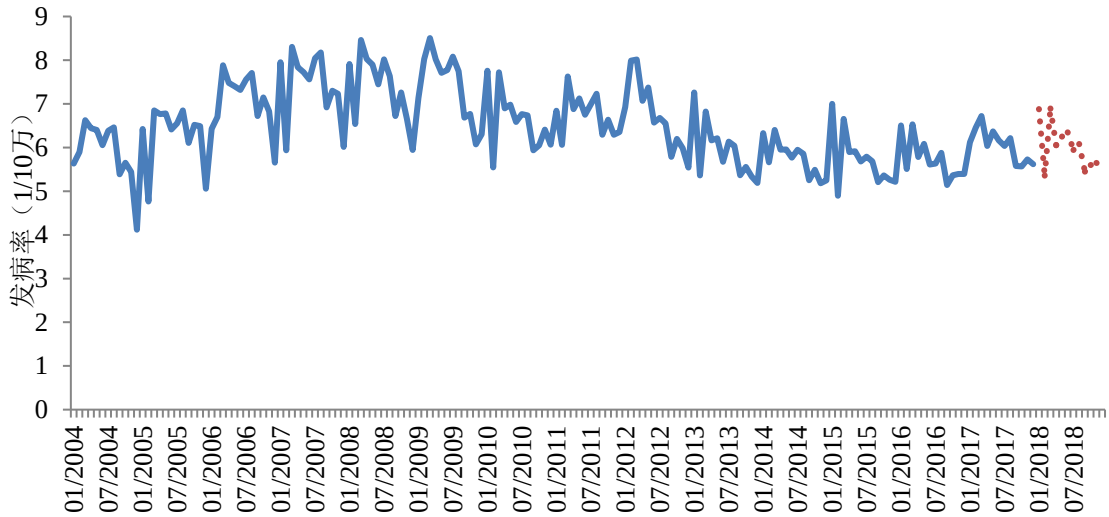


图 4-12　历年发病趋势及 2018 年乙肝发病率变化趋势

2017年实际月发病率最低为5.57/10万，2018年预测发病率最小值为5.45/10万；2018年发病率最高达到6.89/10万，相比2017年最高月发病率6.72/10万有所上升。在整个社会大环境不变的基础下，乙肝发病率的季节趋势不变，发病率略有上升，波动范围相比上一年变大。针对该预测结果，后续全国乙肝防控需重点关注3月份、1月份等高发病月份，减少高发时期的乙肝发病数。

5 结论与展望

5.1 结论及思考

论文通过2004-2017年全国乙肝发病相关数据，应用流行病学方法和空间统计方法展开分析，从传统流行病学特征、时空分布特征和趋势特征预测方法三个层面进行探讨，得到如下基本结论。

5.1.1 结论

（1）从传统三间分布分析得出：第一，时间分布上，我国乙肝发病率在2004年到2006年发病率和死亡率同时上升，2006年死亡率达到最高值，2007-2015一直呈下降趋势，2016年开始，发病率和死亡率略有上升，发病率波动范围在68.57/10万到89.00/10万之间，年均发病率为77.47/10万。乙肝高发病率主要分布在春夏季，1月份、8月份发病率都较高，3月份是一年发病率的最高时期。相比之下，冬季发病率较低。第二，地区分布上，我国乙肝发病率有明显的地区差异，青海、新疆等西北地区乙肝发病率高，随着时间推移，广东和福建地区发病率增加。第三，人群分布上，其中性别分布特征：2004-2017年男性乙肝平均发病率为96.58/10万，女性乙肝平均发病率为57.29/10万。男性乙肝发病数一直高于女性乙肝发病数，男性发病率最高值为111.64/10万，女性发病率最高值为65.49/10万；年龄分布特征：不同年龄段都存在乙肝发病情况，青壮年发病数最高。随着时间变化，低年龄发病数逐渐下降，而40岁以上人群发病数增多。职业分布特征：职业为农民的乙肝发病数最多，其次是从事家政、家务及待业人群，发病数占总病例数的10.46%。

（2）时空相结合方法分析得出，乙肝发病率在空间分布上有明显的高集聚区和低集聚区。从2004-2017年看，我国乙肝发病率的全局Moran's I指数都显著大于0，乙肝发病率存在空间正相关性，表现为高发病率地区与高发病率地区相邻，低发病地区趋于和发病率低的地区相邻。通过局部相关性分析进一步得出，高-高集聚区主要是新疆、青海及周边地区，低-低集聚区为江苏和上海。

（3）从疾病预测角度来看：一方面，常用的单一线性模型和非线性模型，

其预测效果均不占优，就论文比较结果而言，ARIMA 模型的均方误差、平均绝对误差、平均绝对百分比误差和 Theil 不等系数分别为 0.165、0.25、4.04%、0.0337。BP 神经网络模型预测结果显示：均方误差、平均绝对误差、平均绝对百分比误差和 Theil 不等系数分别为 0.075、0.21、3.50%、0.0226。组合模型可以同时考虑单一模型的优点，ARIMA-BP 组合模型既可以提取发病率的线性趋势，又可以提取乙肝发病率的非线性趋势，模型均方误差、平均绝对误差、平均绝对百分比误差和 Theil 不等系数分别为 0.053、0.18、2.95%、0.0190。综合以上指标分析结果，组合模型预测值更接近真实值，其预测精度高，在乙肝发病率预测中更占优。另一方面，根据选取的组合预测模型，发掘未来乙肝发病趋势。结果显示 2018 年乙肝月发病率区间在 5.45/10 万-6.89/10 万之间，发病率的平均水平为 6.01/10 万，标准差是 0.518，季节趋势不变，全国乙肝发病率最高的月份是三月份，冬季发病率较低，与历史实际发病率趋势保持一致。2018 年乙肝发病趋势相对稳定，发病率略有上升，波动范围相比上一年变大。

5.1.2 思考

（1）影响乙肝分布差异原因

第一，2004 年到 2006 年发病率和死亡率同时上升，这可能与我国传染病检测系统发展及传染病报告系统的改变有关，2004 年发病数据采取网络实时录入，部分已患病人群过去可能没有及时统计，导致 2006 年发病率明显上升。乙肝高发病率主要分布在春夏季相比之下，冬季发病率较低。乙肝高发病率分布的时间正好是春节结束，这与有些人可能在节日过后才选择医院就诊有关；此外，在返工过程中，大部分人群的迁移增加了接触性传播传染乙肝的概率[36]。另外，目前的疫情报告数据准确性急需加强，相关研究显示，在对全国范围内医疗机构的抽样调查中，存在乙肝跨年重复报告和在不同医院就诊而重复报告的问题[29]。

第二，我国乙肝发病率有明显的地区差异，青海、新疆等地区乙肝发病率高，随着时间推移，广东和福建地区发病率增加。西北地区发病率高可能和当地经济发展、卫生条件等因素有关，此外，不同地区的环境、饮食习惯都有可能导致地区发病率差异。

第三，乙肝男性发病数和发病率明显高于女性，导致这一现象的原因可能和

男性的社会交际和生活习惯有关，经常在工作应酬时喝酒吸烟会对肝脏产生危害。此外，乙肝主要通过母婴、血液、体液和性接触传播，吸毒、纹身行为中男性居多，这些行为习惯都在很大程度上增加了男性传染乙肝的风险。根据一项普通体检人群的调查显示，女性、本科及以上学历人群接种疫苗必要性的认同率高于男性[23]，这表明男性乙肝率高和该人群对乙肝疫苗的态度有关，后续极需改善男性人群对乙肝疫苗知信行情况。

第四，20-29 岁人群发病数最多，该年龄段人群刚刚接触社会，好奇心比较强，也是参与社会活动的活跃人群，这些原因都有可能导致他们成为乙肝病毒感染的重点人群。此外，公众对乙肝相关知识了解不够完整和全面，认知程度较低。根据对广州市乙肝患者的一项调查显示，乙型肝炎患者对乙肝的一般认知、预防知识掌握程度不高，乙肝科普教育接收率低[12]。许燕[34]等人对成人乙肝疫苗接种情况展开调查，结果表明乙肝发病和高年龄、不了解乙肝相关知识健康教育有一定关系。通过知网和维普等数据库搜集“大学生”、“乙肝”、“认知”相关文献得出，大学生这一群体对乙肝知识没有系统地掌握，专业、生源地这些因素影响着大学生对乙肝的认知[18]。不仅公众对乙肝的了解不够全面，从事医务人员对乙肝的认知也有很大欠缺，在对某高校护理专业学生的调查表明，不同年级学生，实习前后学生对乙肝感染相关知识了解程度低，乙肝的治疗方法知晓程度普遍较低[39]。2002 年起卫生部将乙肝疫苗纳入儿童免疫规划，我国低年龄人群的乙肝发病率得到了有效控制，这与论文研究结果保持一致，从 2004-2017 年，0-19 岁人群发病数逐渐减少，10-19 岁的发病数从 150015 例下降到 19380 例。与此同时，高年龄发病人数逐渐增多，一方面，这部分人在出生时期没有接种疫苗，且后期接种疫苗的免疫失败率高于新生儿；另一方面，老年人对乙肝预防知识不了解，信息获取渠道较窄，这些情况都导致高年龄人群易感染乙肝，并且不能及时发现和治疗。

第五，职业为农民的乙肝发病数最多，农民主要以体力劳动为主，生活环境较差、医疗设备不完善。除此之外，农民文化水平相对不高，不够重视对传染病相关知识的学习，对乙肝的传染性和传播途径没有全面地了解。对于外出打工的农民工而言，在日常工作生活中接触的人员流动性强，这都有可能增加感染乙肝的概率。

（2）乙肝防控对策建议

从不同维度分析全国乙肝分布情况得出，我国新生儿接种乙肝疫苗取得了显著成效，0-19 岁人群发病数明显减少，乙肝的低年龄人群发病情况有很大改善，高年龄人群发病数在逐渐增加。青壮年人群、男性、农民成为目前的发病主体。基于论文研究的乙肝发病流行特征，为了进一步做好乙肝防控措施，降低全国范围的乙肝发病率，应该从以下几点考虑：

第一，首先我们应该从乙肝认知出发，加强关于乙肝的宣传教育工作，重点关注乙肝发病率高的农民、青壮年和男性群体。为了减少高危人群的发病数，积极倡导公众全面体系地了解乙肝相关知识，做好自我防控。其次加强公众对接种乙肝疫苗必要性的认同度，使其积极主动地接种乙肝疫苗，及时进行乙肝血清检测，早预防，早发现，早治疗。最后，对于乙肝高发病地区给予更多的政策支持，根据发病差异，针对不同地区精准施策。不仅要落实好新生儿疫苗接种，提高乙肝疫苗首针及时接种率，而且要根据乙型肝炎的流行特点，对乙型肝炎疫苗免疫程序进行优化调整，增加成人和特殊人群的疫苗接种覆盖率[9]。乙肝发病率高且传染性强，患者疾病治疗的经济压力较大，考虑到我国发病人数中职业为农民的群体占了很高比重，后续对该部分高发病人群要给予更多政策支持并加强乙肝防疫知识教育。此外，我国乙肝发病率在空间分布中存在集聚情况，为了防止相邻省域之间疾病传播，在疾病防治过程中，对乙肝高发病率地区和邻接地区都应做好防控工作，阻断可能的传染源头。

第二，我国法定传染病报告数据在统计过程中应该更加严谨，保证疾病发病数据的真实有效，防止统计过程中出现误差，将最真实的数据呈现出现，便于后续对乙肝发病情况的探索分析，为后续构建合理准确的量化体系奠定基础。

5.2 研究展望

通过分析全国乙肝的发病流行特征并对乙肝的发病数据进行建模拟合，为进一步研究的方向提出以下几点展望。

（1）从获取的乙肝发病率数据来讲，由于我国法定传染病报告目前公开的数据从 2004 年开始，所以论文收集的数据量较少，不能完整地分析乙肝从最初被发现到后续传播的具体发病情况。后期在学习中获取更多乙肝发病率相关数据，

可以做进一步研究。

（2）论文在选取预测乙肝发病率的方法时，从乙肝发病率时间序列特征出发，综合考虑发病序列包含的线性趋势和非线性趋势，构建 ARIMA 模型、BP 神经网络模型和其组合模型。通过对比各模型预测效果，得出组合模型对乙肝发病率预测效果较好。该组合模型没有具体分析不同因素对乙肝发病的影响程度，但在实际生活中，传染病的传播受到很多因素不同程度的影响，比如温度、湿度等气象条件，经济水平，医疗卫生设备，人口流动量等。所以在后续发病预测研究中，如果充分结合乙肝的具体传播途径，将影响乙肝发病及传播的可能因素引入模型，可使其预测模型更准确完善。

（3）目前我国乙肝疾病的防治取得显著成效，乙肝发病率有了明显下降，但由于我国人口数多，现有的乙肝发病数仍居高不下。为减少我国乙肝发病数，在后续乙肝疾病的防控和监管中，我们不但要切实遵循“预防为主，防治结合”的原则，而且要重点关注乙肝高发病群体，采取调查等方式研究乙肝发病率高的具体原因，以便针对性地采取防控措施并及时布局。此外，由于我国目前成人乙肝患者数较多，除了要做好乙肝防控，还应该关注实际治疗。患者在治疗乙肝时需承担巨大经济压力，为进一步减少乙肝发病数、降低乙肝发病率，后续研究应关注国家在乙肝防治中投入的资金支持，切实监查医疗产品，从多个角度进行量化分析，构建系统全面的量化体系，最终为乙肝防治提供理论依据。

参考文献

[1]EmmaK,Chaput JIM,Robert H.Spatial analysis of human granulocytie ehrlichiosis near lyme,Connecticut[J].Emerging Infectious Diseases,2002,8:943-948.

[2]Kohonen,T(1988).An introduction to neural computing.Neural Networks, 1(1):3-16.

[3]Map-making and myth-making in Broad Street: the London cholera epidemic, 1854[J].Howard Brody,Michael Russell Rip,Peter Vinten-Johansen,Nigel Paneth, Stephen Rachman.The Lancet.2000,356(9223):64-68.

[4]McCulloch,W.S. and W. Pitts. (1943).A logical calculus of the ideas immanent in nervous activity.Bulletin of Mathematical Biophysics,5(4):115-133.

[5]Moran PAP.The interpretation of statistical maps[J].Journal of the Royal Statistical Society.Series B (Methodological),1948,10:243-251.

[6]Riedmiller,M.(1994) Rprop—Description and Implementation Details.Technical Report,University of Karlsruhe,Karlsruhe.

[7]Rui Yu,Rong Fan,Jinlin Hou.Chronic Hepatitis B virus infection:epidemiology, prevention,and treatment in China[J].Frontiers of Medicine,2014,8(2):135-144.

[8]Shahdoust Maryam,Sadeghifar Majid,Poorolajal Jalal,et al.Predicting Hepatitis B monthly incidence rates using weighted Markov chains and time series methods. 2015,15(1):28-31.

[9]Xueyan Liao,Zhenglun Liang.Strategy vaccination against Hepatitis B in China[J]. Human Vaccines & Immunotherapeutics,2015,11(6):1534-1539.

[10]Ya-wen Wang,Zhong-zhou Shen,Yu Jiang.Comparison of ARIMA and GM(1,1) models for prediction of Hepatitis B in China[J]. PLOS ONE, 2018,13(9).

[11]Yuliang Xi,Fu Ren,Shi Liang,Jinghua Zhang,De-Nan Lin. Spatial Analysis of the Distribution,Risk Factors and Access to Medical Resources of Patients with Hepatitis B in Shenzhen,China[J].IJERPH,2014,11(11):11505-11527.

[12]陈敏儿.49 例乙型肝炎患者对乙肝疾病认知和预防情况的调查分析[J].智慧健康,2020,6(23):187-190.

[13]陈强.高级计量经济学及 Stata 应用[M].高等教育出版社,2014.

[14]陈婷.ARIMA 模型和 BP 神经网络模型在艾滋病发病率预测应用中的比较研究[D].广西医科大学,2015.

[15]崔富强,王富珍,吴振华,龚晓红,陈园生,郑徽,缪宁.中国 2005～2010 年报告乙型病毒性肝炎发病分析[J].中国疫苗和免疫,2011,17(06):483-486+559.

[16]邓秋云,钟革,刘巍,韦敬航,杨仁聪,杜进发,董爱虎,黄影.2018 年广西壮族自治区 1-59 岁人群乙型肝炎血清流行率调查[J].中国疫苗和免疫,2020,26(01):25-29.

[17]范珂,张红莲.328 例住院乙肝患者乙肝防治知信行调查及影响因素分析[J].河南预防医学杂志,2020,31(04):280-282.

[18]付之鸥,周扬,陈诚,郑洪伟,宋伟,李苑,陆伟,彭志行.时间序列分析与机器学习方法在预测肺结核发病趋势中的应用[J].中国卫生统计,2020,37(02):190-195.

[19]黄凤.我国大学生乙肝认知调查研究进展[J].科教导刊(上旬刊),2020(06):190-192.

[20]姜庆五,赵飞.空间自相关分析方法在流行病学中的应用[J].中华流行病学杂志,2011,032(006):539-546.

[21]李兰娟,任红主编.传染病学(第 8 版)[M].北京:中国人民卫生出版社,2013.

[22]李立明.流行病学[M].北京:人民卫生出版社,2007.

[23]刘观秀.500 名普通体检人群乙肝疫苗知信行调查及护理建议[J].医学理论与实践,2020,33(01):147-148.

[24]刘琼,杨建华.隐马尔科夫模型在乙肝发病预测中的应用[J].数学的实践与认识,2017,47(19):203-210.

[25]乔贺倩.应用智能组合模型预测中国肺结核月发病人数[D].兰州大学,2018.

[26]覃柳麻.2012-2017 年南宁市乙肝流行特征及空间分析[D].广西医科大学,2019.

[27]史雯,周洋,袁辰,严睿,唐学雯,何寒青,邓璇.2018 年浙江省健康人群乙型肝炎血清流行病学调查分析[J/OL].疾病监测:1-7.

[28]王劲峰,廖一兰,刘鑫.空间数据分析教程[M].北京:科学出版社,2010.2：101-102.

[29]王丽萍,郭青,张春曦,郭岩,赵自雄,马家奇,杨功焕.2006 年全国乙型病毒性肝炎报告质量调查分析[J].中华疾病控制杂志,2009,13(01):69-71+102.

[30]王平.三种预测模型在主要传染病发病率预测中的应用[D].浙江大学,2010.

[31]王燕.应用时间序列分析[M].中国人民大学出版社,2008.

[32]肖占沛,路明霞,张明瑜,张肖肖,马雅婷,王长双,王燕,张延炀.河南省 2009-2019 年风疹流行特征和时空聚集性[J/OL].中国疫苗和免疫:1-8.

[33]谢骁旭.基于R的江西省肺结核发病率ARIMA-SVM组合预测模型[D].南昌大学,2015.

[34]许燕,吴青青,徐水洋,徐锦杭,黄玉.浙江省部分地区成人乙肝疫苗接种情况及影响因素分析[J].中国健康教育,2020,36(03):259-261+284.

[35]杨品超,张顺祥,孙盼盼,蔡亚丽,林莹,邹宇华.乙型肝炎防治经济学评价——马尔科夫模型的构建[J].中华流行病学杂志,2017,38(07):845-851.

[36]杨晓丽.2007-2016 年辽宁省乙型病毒性肝炎流行特征及预测[D].吉林大学,2019.

[37]于颖慧.我国手足口病流行特征及预测模型的建立[D].吉林大学,2019.

[38]瞿嵘.基于空间数据分析技术的流行病空间格局研究[D].武汉大学,2013.

[39]赵志莹,徐胜波.护生对乙肝认知、信念及行为情况的调查分析[J].世界最新医学信息文摘,2019,19(50):293-295.

[40]张靳冬,张建陶,钱建东,潘明珠.灰色系统 GM(1,1)模型在常州市乙型肝炎发病趋势预测中的应用[J].蚌埠医学院学报,2013,38(04):476-478.

[41]周红霞,唐咸艳,仇小强.空间流行病学理论与方法研究现状与展望[J].国外医学:医学地理分册,2015(2):79-92.

[42]周强,孙传武,毕俊.灰色系统 GM(1,1)模型在徐州市乙型肝炎发病趋势预测中的应用[J].职业与健康,2016,32(24):3435-3437.

[43]周志华.机器学习及其应用[M].清华大学出版社,2009.

[44]钟少波.GIS和遥感应用于传染病流行病学研究[D].中国科学院研究生院(遥感应用研究所),2006.

附 录

附表 1 2004-2010 年乙肝月发病率（1/10 万）

月份	2004	2005	2006	2007	2008	2009	2010
1 月	5.6336	6.4234	6.4266	7.9525	7.9138	7.1355	7.7584
2 月	5.8957	4.7634	6.6980	5.9358	6.5391	8.0241	5.5460
3 月	6.6262	6.8503	7.8838	8.3047	8.4637	8.5060	7.7225
4 月	6.4431	6.7639	7.4781	7.8389	8.0255	8.0184	6.8971
5 月	6.4034	6.7796	7.4023	7.7263	7.8975	7.7090	6.9819
6 月	6.0530	6.4123	7.3177	7.5606	7.4487	7.7744	6.5874
7 月	6.3828	6.5521	7.5570	8.0439	8.0173	8.0806	6.7646
8 月	6.4596	6.8501	7.7074	8.1758	7.6321	7.7477	6.7352
9 月	5.3893	6.1070	6.7223	6.9183	6.7218	6.6850	5.9378
10 月	5.6530	6.5186	7.1467	7.3018	7.2608	6.7680	6.0520
11 月	5.4419	6.4900	6.8283	7.2315	6.6537	6.0731	6.4117
12 月	4.1191	5.0575	5.6564	6.0148	5.9432	6.3027	6.0652

附表 2 2011-2017 年乙肝月发病率（1/10 万）

月份	2011	2012	2013	2014	2015	2016	2017
1 月	6.8434	6.9272	7.2592	6.3295	6.9990	6.5055	6.1188
2 月	6.0602	7.9891	5.3624	5.6626	4.8976	5.5074	6.4525
3 月	7.6269	8.0146	6.8237	6.4019	6.6548	6.5295	6.7222
4 月	6.8797	7.0682	6.1644	5.9517	5.8968	5.7849	6.0373
5 月	7.1237	7.3721	6.2137	5.9567	5.9145	6.0826	6.3714
6 月	6.7515	6.5699	5.6729	5.7696	5.6801	5.6106	6.1655
7 月	6.9879	6.6788	6.1366	5.9477	5.7923	5.6314	6.0372
8 月	7.2299	6.5530	6.0412	5.8549	5.6827	5.8785	6.2159
9 月	6.2909	5.7860	5.3640	5.2509	5.2085	5.1407	5.5796
10 月	6.6368	6.1968	5.5549	5.4896	5.3612	5.3671	5.5678
11 月	6.2909	5.9864	5.3353	5.1822	5.2643	5.3954	5.7284
12 月	6.3567	5.5412	5.1902	5.2495	5.2161	5.3954	5.6170

致 谢

时光飞逝，三年硕士生涯已经接近尾声，内心充满感恩与不舍，还清楚记得考研面试时的紧张心情，现在回想起来仿佛如昨日。感恩，首先感谢我亲爱的赵煜老师。本科阶段就上过您的课，从那时就很喜欢您的教学风格。本科阶段没有真正接触过各种项目，在硕士研究生学习期间赵煜老师给了我很多学习的机会，通过参与项目，跟老师的交流，老师教会我严谨的工作和科研态度。不论是在校期间还是外出实习期间，老师都关心着我的学习和生活。在项目中遇到问题时老师总能找到突破口使项目顺利推进，我也学会了遇到问题时不能慌，要多思考，想办法解决问题。在我遇到技术难题时，老师总能给我提供一些新的思路，从不同角度考虑问题，让我大胆的尝试。在写此论文的整个过程中，赵老师始终细心指导，即使在身体不适的情况下，也指导我发现论文中的问题并进一步完善，再次向给予我无私帮助的老师致以最诚挚的感谢！感谢赵煜老师这三年来对我的影响和谆谆教诲。祝老师身体健康、未来生活一切如意。

同时感谢室友和朋友，在论文完成过程中感谢他们的鼓励和帮助；感谢遇见彼此，让生活变得多姿多彩。未来的你们，一定会过上自己想要的生活。

在此还要感谢我的家人，在 20 多年的学习生涯中，您始终支持我、信任我，为了我的成长付出汗水和泪水。希望在以后的日子里，您可以无忧无虑，为您自己而活。

最后，感谢各位评审老师，您辛苦了。

至此，怀念我的硕士研究生时期。

毛少霞

2021 年 4 月

兰州财经大学