所属类别	2024 年第三届全国大学生数学分析实践赛	参赛编号
本科组		2408228

基于时间序列分析方法的中国医疗卫生发展分析

摘要

近年来，随着脱贫攻坚战的全面胜利，我国医疗卫生事业也取得了瞩目的成绩，但是地区、城乡间医疗卫生资源分配不均衡的问题依然存在；脱贫解决了居民“衣食住行”的问题，但并没有真正解决一些地区人们“看病难”的忧虑。为了进一步了解我国各地区医疗卫生发展水平，分析医疗卫生资源配置的影响因素，为未来我国医疗卫生事业发展建言献策，本文针对 1990 年至 2023 年我国及地区医疗卫生相关数据，使用 Python 研究过去 33 年各地区医疗卫生事业发展的趋势及差异，并采用 Holt-Winters 模型、Seasonal AutoRegressive Integrated Moving Average（SARIMA）模型和与线性回归的组合模型对中国 31 个地区未来 5 年医疗卫生机构数（以下简称“机构数”）进行预测，并采用 Long Short-Term Memory（LSTM）模型宏观预测未来 10 年中国机构数。

针对问题一，本题需要计算各地区机构数的年均增长率并对比它们的增长趋势。本文使用复合年增长率（CAGR）来衡量地区机构数的增长趋势，并绘制折线图来可视化各地区增长趋势，绘制柱状图比较地区年增长率的大小，结果显示河南、河北、山东三省增长最快，而上海却呈现负增长的趋势。

针对问题二，本题需要研究各地区医疗卫生资源配置的差异，并分析其可能的影响因素，探讨结果背后的原因。本文从人力、物力、财力三个角度出发，选择 6 个指标来评价地区医疗卫生资源配置的水平；之后针对每个指标，采用箱线图对比地区数据的整体分布差异，采用柱状图对比地区数据的年均增长率，采用热力地图可视化地区年均增长率的地理差异，最后用综合评价方法，计算各地区医疗卫生资源配置的综合得分，得分区间为[0, 1]对，以比它们的配置差异。结果显示，辽宁省综合资源配置水平最高，海南省的水平最低，但两者之间得分差异仅为 0.19。

针对问题三，本题需要预测各地区未来 5 年（2024 年至 2028 年）的机构数，并讨论预测结果。本文将数据集中 31 个地区根据其历史机构数发展模式划分为 3 类，并针对 3 类地区分别使用 Holt-Winters 模型、SARIMA 模型和前两者与线性回归的组合模型进行机构数预测，模型检验结果显示，所有地区预测值的R^2指标均大于 0.1，且从可视化结果来看，模型预测值基本符合数据变化趋势，具有可靠性。

针对问题四，本题需要从多因素分析角度，评估影响机构数的关键因素，并宏观预测未来 10 年（2024 年至 2033 年）中国的机构数。本文从医疗卫生资源配置、人口结构、经济发展水平、政策四个角度出发，选择 10 个因素进行影响因素分析及预测；通过皮尔逊相关系数分析，发现人口密度与机构数几乎不存在线性相关性，做删除处理；之后使用随机森林回归模型对剩下 9 个因素做重要性排序，结果显示，卫生人员数对机构数影响最显著；最后应用 LSTM 模型做多变量预测，发现未来 10 年中国机构数呈缓慢的波动下降趋势。

关键词：医疗卫生机构数，Holt-Winters 模型，SARIMA 模型，LSTM 模型

1. 问题背景与重述

1.1 问题背景

随着我国经济和综合实力的提升，公共卫生领域投入不断加大，医疗科技水平迅速提高，医疗事业取得了显著进展。全国卫生医疗机构和卫生技术人员增长，但是医疗资源紧张且地区、城乡不均衡问题日渐显著，优质的医疗资源主要既注重在经济发达地区。对于资源匮乏或是不平衡地区，没有足够的资源满足人民的看病需求，影响百姓的生活质量。

此背景下，了解我国各地区医疗卫生发展水平，如医疗卫生机构数量——衡量医疗卫生资源配置的重要指标，且受政策与经济发展水平等因素影响，分析不同时期和不同地区医疗卫生机构数量和变化趋势，了解类似于医疗卫生机构数量等指标及其影响因素，了解我国及各地区医疗卫生大致水平，对于合理配置医疗资源、完善医疗制度体系至关重要。

1.2 问题重述

问题一，医疗卫生机构数量数据处理与分析。问题一分为两个部分，年均增长率与时间趋势分析和数据可视化比较。前者先处理已给数据，并计算医疗卫生机构数的年均增长率进行初步比较，后者则将数据可视化，绘制图表比较不同地区的差异。

问题二，医疗卫生资源配置差异及影响因素分析。基于问题一的分析结果且结合其他相关数据讨论中国各地区医疗卫生资源配置的差异及其可能的影响因素。揭示医疗卫生机构数量的区域差异和变化趋势，并探讨背后的原因及政策意义。

问题三，各地区未来 5 年医疗卫生机构数发展预测。利用 1990 年至 2023 年的数据，预测未来五年各地区医疗卫生机构数量，并且评估模型的准确性，讨论预测结果的可靠性和潜在的不确定性因素，并分析这些因素对未来医疗资源配置的影响。

问题四，中国未来 10 年医疗卫生机构数宏观预测与政策启示讨论。基于医疗资源配置、人口结构变化、经济发展水平和政策导向等因素对中国未来十年的医疗卫生机构进行宏观预测，并讨论对政策制定、资源分配和体系建设的启示、挑战和应对策略。

2. 问题分析

2.1 问题一分析

第一部分需要计算医疗卫生机构的年均增长率，根据已提供的各地区医疗卫生机构数量的时间序列数据先进行填补。首先查看有缺失值的省份及缺失个数，缺失数较少，且指标数据为时间序列数据，具有趋势性和自相关性，所以可假定缺失部分数据随年份的自变量变化而呈局部线性变化。在用线性插值方法填充数据之前，先采用三次样条插值方法填充数据，并检查数据是否填充完成，发现模型拟合效果欠佳。于是使用效果更好的线性插值方法，填充完毕后，利用折线图比较不同地区医疗机构变化情况。计算年均增长率函数并进行时间趋势分析，识别各地区医疗资源的增长速度与模式，确定增长最快和增长较为缓慢的地区。第二部分则使用前一部分的计算结果进行可视化操作，绘制柱状图来更直观地分析各地区的机构数的变化差异。

2.2 问题二分析

根据所搜资料，本文主要以人力、物力与财力三个方面评价医疗卫生资源的配置，涉及的主要指标包括卫生技术人员数、职业（助理）医师人员数、注册护士数、卫生机构床位数、医疗卫生机构数、财政支出中卫生经费等可能影响医疗资源配置的因素。接

着在对每个方面的指标在分析之前，将每个参数的缺失数据预处理，进行填充。接着将这三个方面的每个参数使用箱线图、平均增长率柱状图与平均增长率热力图来进行数据可视化，三个图分别从数据分异程度、时间变化趋势和地理特征来比较各省份每个指标的差异。接着将数据进行标准化处理，分配比重进行综合评价，并使用热力图与柱状图来进行可视化对比。最后观察各地区在不同参数的变化趋势与差异，并讨论其成因。

2.3 问题三分析

预测各地区 5 年的医疗卫生机构数量变化，因为可使用因素数据较少，且可使用数据时间跨度不一致且较短，为了减少因为其他数据造成的误差，所以本文采用单因素数据预测，即用 1990 至 2023 年机构数量预测未来 5 年机构数量。参考第一问的机构数随时间变化图，可将各地区数据分为三类——呈明显上升趋势且又小幅度波动的、呈现剧烈波动的及呈现较大波动且有上升趋势的数据。

本文在预测第一类数据时，使用 Holt-Winters 指数平滑模型处理具有趋势性的时间序列数据，将数据分为训练集和测试集，并使用 MAE 为评价指标来动态调整预测值，适应数据的变化。预测第二类数据时，通过季节性和非季节性的结合，SARIMA 模型能够捕捉数据中的复杂波动特征。预测第三类则组合使用这两种方法。

2.4 问题四分析

本文在考虑影响医疗卫生机构数量的关键因素时，列出了医疗卫生资源配置、人口结构变化、经济发展水平和政策导向的指标，如卫生人员数及医药制造业专利申请数来衡量医疗卫生资源配置。接着预处理数据，使用线性插值填充这些指标的缺失数据。在进一步分析时使用 seaborn 绘制热力图来可视化相关关系矩阵计算不同指标与医疗卫生机构数的相关性。在删去几乎没有相关性的指标后进一步使用随机森林来分析各个因素的重要性。首先要对特征数据进行标准化，将数据转换为均值为 0、方差为 1 的标准正态分布。在显示各个因素的重要性后预测未来十年的医疗机构数时，设置随机种子以保证结果的可复现性。

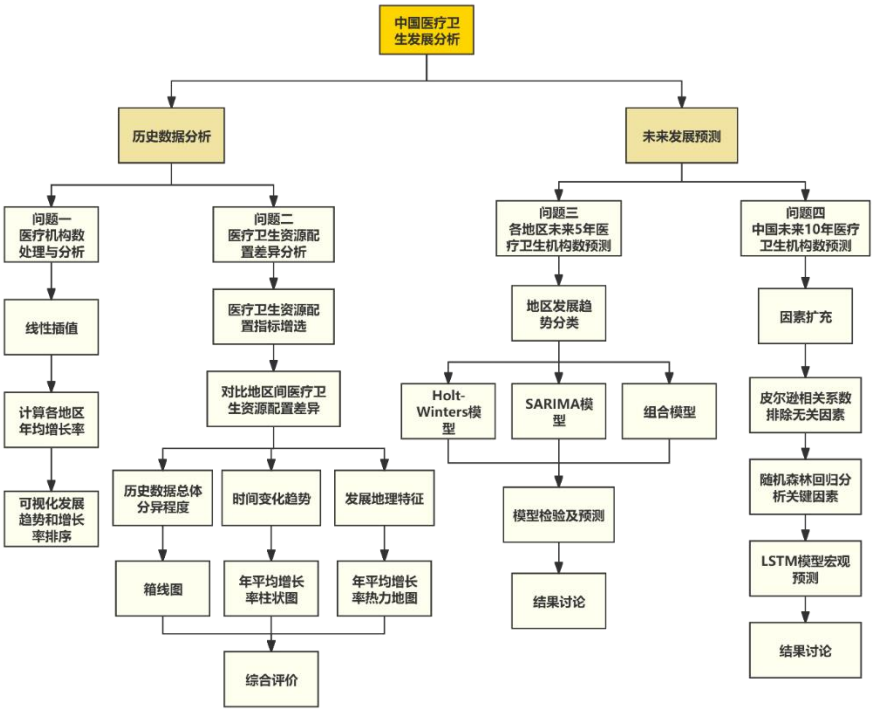


图 2-1：本文分析思路

3. 模型假设

假设一：在地区数据缺失的年份数量合理的情况下，即缺失数据年份不多于统计总年份20%的情况下，假设缺失年份的数据与之后的第一年没有缺失的数据呈线性关系。

假设二：假设已统计且没有缺失的指标数据均真实可靠，即排除异常值的可能，可以直接用于数据分析和预测。

假设三：假设模型对历史数据的预测误差和对未来数据的预测误差属于同分布，可以用历史预测误差来校正模型对未来数据的预测值。

4. 符号说明

变量	说明	量纲
t	年份	个
y_t	在t年份的真实数据	/
p	SARIMA 自回归阶数	/
d	SARIMA 差分阶数	/
q	SARIMA 移动平均阶数	/
$RMSE$	均方根误差	/
MAE	平均绝对误差	/
R^2	模型拟合优度的统计量	/
$CAGR$	复合年均增长率	%
lr	梯度下降的学习率	/
$epochs$	LSTM 模型训练轮数	轮
$hidden$	LSTM 模型隐藏层数量	层
$layers$	LSTM 模型循环层数量	层

5. 问题一模型建立与求解

5.1 数据预处理

5.1.1 缺失数据填补

5.1.1.1 线性插值填充模型建立

提供的 1990 年至 2023 年的各省医疗机构数量的数据中，大约 45%的省份有不同程度的缺失，缺失数据有 28 个，大约占总数据的 3%，具体缺失数量如表 5-1：

表 5-1：医疗卫生机构数据缺失地区及个数

地区	缺失个数
天津市	2
山西省	1
内蒙古	1
辽宁省	2
吉林省	2
黑龙江	1
上海市	2
江苏省	1

浙江省	2
湖南省	2
海南省	2
重庆市	6
西藏	2
陕西省	2

且医疗机构数量的数据属于时间序列数据，具有以下特征：

1. 时间依赖性。医疗机构数量数据随着时间的推移而记录，具有强烈的时间依赖性。过去某一时刻的医疗机构数量会影响未来的数量变化。

2. 趋势性。医疗卫生机构可能表现出长期的上升或下降趋势。

3. 季节性。但在某些情况下，例如每年特定时期的政策变化或预算分配，可能会影响短期内的机构数量波动。

4. 波动性。同年份之间，医疗机构数量的增长可能会受到经济环境、政策导向等因素的影响，表现出一定的波动性，但通常不会有剧烈的短期波动。

5. 自相关性。医疗机构数量可能与前几年的数量高度相关，反映出数量变化的惯性和延续性。

本文使用线性插值模型填补缺失数据。在填补缺失值之前，通过绘制 1990 年至 2023 年间各省份的医疗卫生机构数量变化趋势图[1]，直观地观察数据的变化情况。此步骤帮助确认数据的趋势是否符合线性插值的假设，即数据在时间轴上应表现出相对平稳和线性的变化趋势。使用线性插值方法假设有一组数据点使用线性插值方法假设有一组数据点(x_1,y_1) 和(x_2,y_2) ，其中x_1和x_2是是已知点的横坐标（年份），y_1和 y_2是已知点的纵坐标（对应年份的医疗卫生机构数）。对于位于x_1和x_2之间的某个缺失点 x，其对应的插值结果 y 可以用以下公式计算：

$$y = y_1 + \frac{(x - x_1) \times (y_2 - y_1)}{x_2 - x_1} \qquad (5-1)$$

医疗卫生机构数量表现一定的趋势性，且通常没有剧烈的短期波动，使用线性插值模型可以较好地捕捉这种趋势，且避免了像高阶多项式或样条插值那样，可能引入不必要的波动或“过拟合”现象，从而导致不合理的估计。

5.1.1.2 插值模型求解

使用内置函数线性插值具体流程如下：

1. 数据预处理。将所有缺失值表示符号“--”替换为 NaN，数据框中的缺失值能够被识别，将数据类型转换成数值型，以便后续进行数值计算和插值。
2. 缺失值分析。识别省份缺失数据的年份，确认缺失值个数。
3. 数据趋势分析。绘制 1990-2023 年各省份的医疗卫生机构数量变化趋势图，直观观察数据的变化情况。
4. 调用函数插值。通过调用 interpolate(method='linear', limit_direction='both') 函数，对整个数据集进行线性插值处理。此方法会对每一个有缺失值的时间点，根据前后已知数据点的值进行线性估算，从而填补缺失值。
5. 插值效果检查。在完成插值操作后，再次检查数据集中的 NaN 值，确保所有缺失值都已被成功填补。

[1] 见附录 2

5.2 计算与分析各地区的年均增长率

5.2.1 计算与绘制年均增长率

本文使用年均增长率（CAGR，Compound Annual Growth Rate）的计算公式计算各地区的年均增长率，计算公式如下：

$$CAGR = \left(\frac{End\ Value}{Start\ Value}\right)^{\frac{1}{n}} - 1 \qquad (5-2)$$

其中，End Value 是期末的数值，Start Value 是期初的数值，n 是计算期间的年数。之后绘制柱状图比较各地区的增长率将数据可视化，如图 5-1：

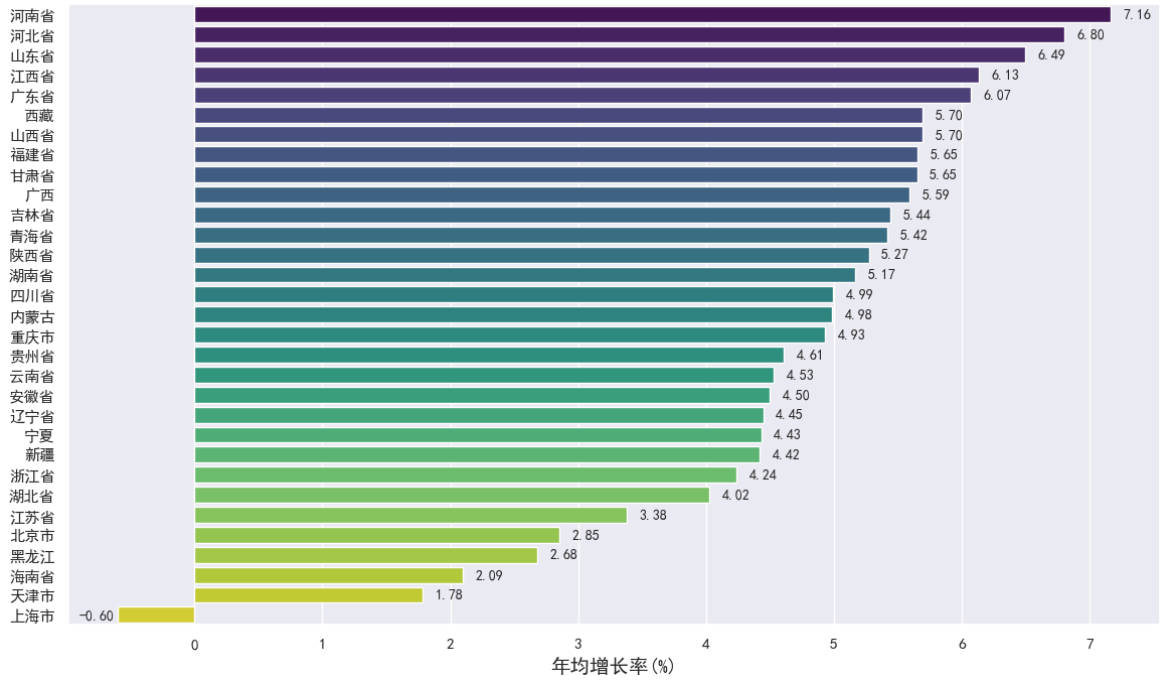


图 5-1：各地区医疗卫生机构数量年均增长率

5.2.2 各地区医疗机构数量年均增长率分析

根据所绘柱状图可以很清楚地看到各地区医疗卫生机构数的年均增长率呈现出来不同的增长趋势，其中河南、河北及广东等省份增长较快，且超过 6%，医疗卫生机构快速增加。初步分析可能与这些省份较高的人口密集程度、经济发展水平及医疗需求增长有关。上海、天津等省份年均增长率较低，上海甚至出现了负增长。初步分析可能因为这些省份地卫生机构数量趋于饱和，或是政策、人口流动因素导致机构数量减少。

总体来看中国的医疗卫生机构数量总体趋势增长，且快速增长的地区集中在中部和东部经济相对发达、人口较多的省份。增长较慢的地区则多为发达地区或直辖市，如上海、北京和天津等。此外，一些边远地区如海南省的增长也相对较慢，如海南。

6. 问题二模型建立与求解

6.1 指标增选

本文除了已给定的医疗卫生机构数量外，考虑到评价医疗卫生资源配置均衡程度还受到其他指标来衡量，于是采用医疗卫生资源配置均衡程度的评价指标体系错误!未找到引用源。

来评价各省份的医疗资源均衡程度，如表 6-1：

表 6-1：医疗卫生机构资源配置均衡程度的评价指标体系

一级指标	二级指标	三级指标
医院和基层医疗卫生机构的医疗卫生资源配置	医疗卫生人力资源	卫生技术人员数 执业（助理）医师人员数 注册护士数
	医疗卫生物力资源	卫生机构床位数 医疗卫生机构数
	医疗卫生财力资源	财政支出中卫生经费

6.2 指标比较

针对每个指标，本文使用箱线图来比较数据分异程度，使用平均增长率柱状图比较时间变化趋势，使用平均增长率的热力地图来比较地理特征。

6.2.1 绘制图表模型求解

绘制各指标各方面箱线图、平均增长率柱状图与平均增长率的热力地图具体流程如下：

1. 加载数据。从 CSV 文件中加载多个指标的数据集。
2. 使用箱线图函数绘制箱线图。将数据按照中位数从高到低排列，转化成成长格式后绘制箱线图。
3. 使用年均增长率函数绘制平均增长率柱状图。计算各地区的年均增长率后将增长率词典封装成 dataframe 后绘制柱状图比较。
4. 绘制热力地图。处理已读入的地图数据，将原数据没有的地区删除，将各省平均增长率数据映射在地图上。

6.3 数据分析

6.3.1 卫生技术人员数差异及其可能影响因素分析

通过箱线图[2]可以观察到以下几个关键点：

1. **广东省**的中位数最高，表明该省在研究期间内卫生技术人员的数量较多，且分布较为广泛。**山东省、河南省、四川省、湖北省**等省份的中位数也相对较高。这些省份大多是经济较为发达且人口较多的省份，卫生资源的配置较为充足，需求较高，且经济发达地区较能吸引人才[0]，说明经济发展水平与人口密度可能是其发展因素。
2. 在一些省份，如**辽宁省**和**吉林省**，可以看到一些离散的点（即异常值），这些异常值可能代表某些年份卫生技术人员的数量显著偏离正常水平，可能由于某些特殊政策、事件或统计异常引起。如辽宁省在 2015-2020 年期间实施了《辽宁省医疗卫生服务体系规划》，这可能对卫生技术人员的数量产生了影响[1]。

[2] 卫生技术人员箱线图见附录 1-1

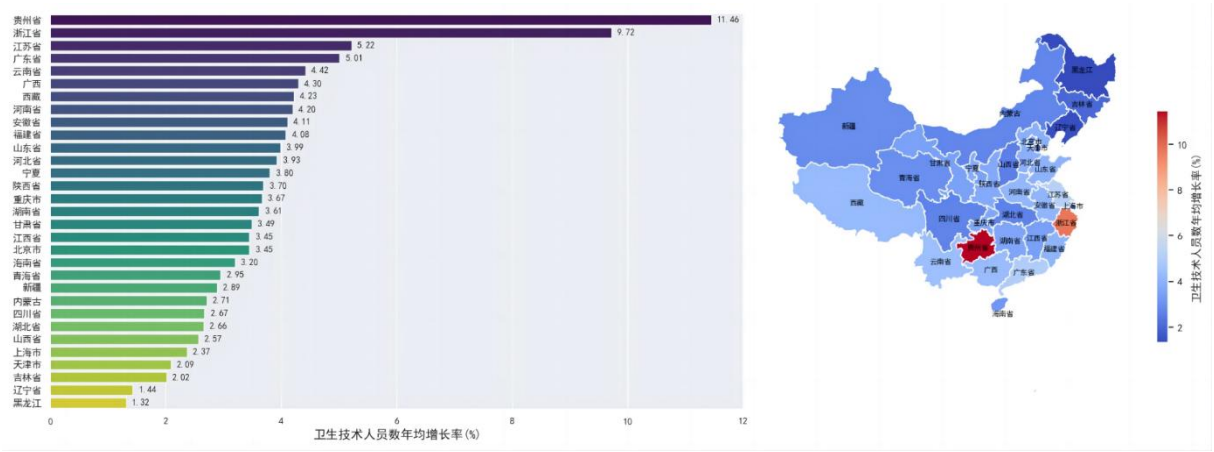


图 6-1：各地区卫生技术人员数年均增长率可视化

图 6-1（左）展示了各省份卫生技术人员数量的年均增长率，图 6-1（右）平均增长率的热力地图反应其地理特征可以观察到几个关键点及其影响因素分析：

1. 贵州省和浙江省的卫生技术人员数量年均增长率最高，分别达到了 11.46%和 9.72%。经济发达的浙江省增长率较高的原因可能是，其在持续增加医疗卫生领域的投资，推动了技术人员的增长。而贵州作为经济发展水平较弱的西部省份，则可能是因为近年来在国家政策的扶持下，医疗资源配置有所加快，如贵州省已出台 58 项支持政策，涵盖政府投入、人才引进等[3]。
2. 黑龙江和辽宁等省的增长率则较低，分别是 1.32%和 1.44%。北京与上海等经济大城市增长率也较低。可能与各地区经济发展策略和人口变化有关。
3. 卫生技术人员数量增长较快的地区主要集中在中西部及东南沿海的几个省份，而东北地区和部分西北地区的增长率相对较低。这表明中西部地区在医疗卫生资源方面正逐步追赶东部沿海发达地区，显示出国家在政策上对中西部地区的扶持成效显著[4]。

6.2.2 执业（助理）医师人员数差异及其可能影响因素分析

通过箱线图[3]可以观察到以下几个关键点：

1. 山东省中位数最高，广东省、江苏省与河南省等省份中位数也较高，表明该省在研究期间内执业（助理）医师人员数较多，且分布较为广泛。这可能是由于山东省人口基数大，对医疗服务的需求较高，从而推动了执业医师数量的增加。广东省、江苏省与河南省等这些省份的中位数也处于较高水平，反映出其执业（助理）医师数量在全国范围内较多。这与这些省份经济发达、医疗资源丰富以及人口密集有关。
2. 在一些省份，如吉林省与上海市，可以看到一些较为密集的离散的点（即异常值），这些异常值可能代表某些年份执业（助理）医师人员数显著偏离正常水平。这可能由于政策变化、重大公共卫生事件，如新冠疫情对全球医疗系统造成的冲击[5]，或经济波动等原因导致的。

[3] 执业（助理）医师人员数箱线图见附录 1-2

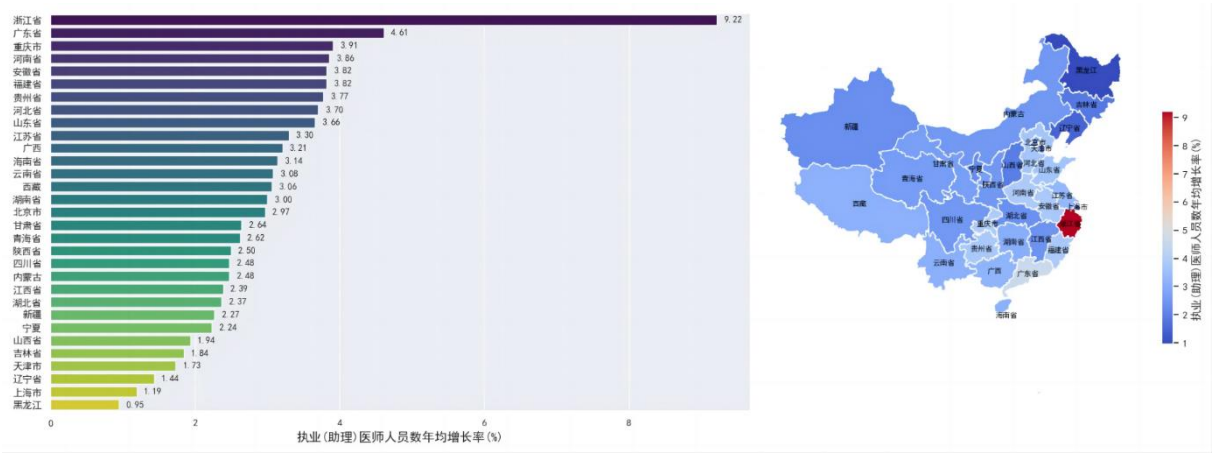


图 6-2：各地区执业（助理）医师人员数年均增长率可视化

图 6-2(左)展示了各省份执业（助理）医师人员数的年均增长率，该图 6-2(右)平均增长率的热力地图反映其地理特征可以观察到几个关键点及其影响因素分析：

1. 浙江省年均增长率最高，浙江省的年均增长率几乎是第二高广东省的两倍，这可能与浙江省近年来医疗卫生改革的推进力度有关，如《浙江省医疗卫生服务体系暨医疗机构设置“十四五”规划》**错误!未找到引用源。**特别是在人才引进、教育培训和医疗卫生设施建设方面的投资增加。
2. 黑龙江省及上海市等年均增长率较低，黑龙江增长率甚至小于 1%。可能的原因包括人口老龄化、经济增长缓慢以及城市吸引力下降等。上海市虽然是一个发达地区，但由于其基础较高，增长空间相对有限，加上医疗卫生人员供求已经趋于平衡，所以增长率相对较低。
3. 执业（助理）医师人员数增长较快的地区主要集中在东南沿海的几个省份，而东北地区和部分西北地区的增长率相对较低。东南沿海省份如浙江、广东等地经济发达，医疗需求增加，同时也拥有更多的资源投入到医疗卫生体系中。这些地区的经济活力和人口增长带来了对医疗服务需求的上升，进而促进了执业（助理）医师人员数的增长。西北这些地区可能面临人口流失、经济发展滞后和医疗资源不足等问题，从而导致执业（助理）医师人员增长的压力较大。此外，这些地区的气候条件、生活环境以及对专业人才的吸引力不足，也可能影响了医师数量的增长。

6.3.3 注册护士数差异及其可能影响因素分析

通过箱线图[4]可以观察到以下几个关键点及其影响因素分析：

1. 广东省中位数最高，山东省、江苏省等省份也较高。经济较为发达、人口密集以及较高的医疗资源配置可能是这些省份护士数量较多的原因。
2. 西藏、青海省与宁夏等省份中位数较低。可能受到地理位置偏远、人口较少以及经济发展水平较低等因素的影响，导致医疗资源相对短缺。

[4] 注册护士数箱线图见附录 1-3

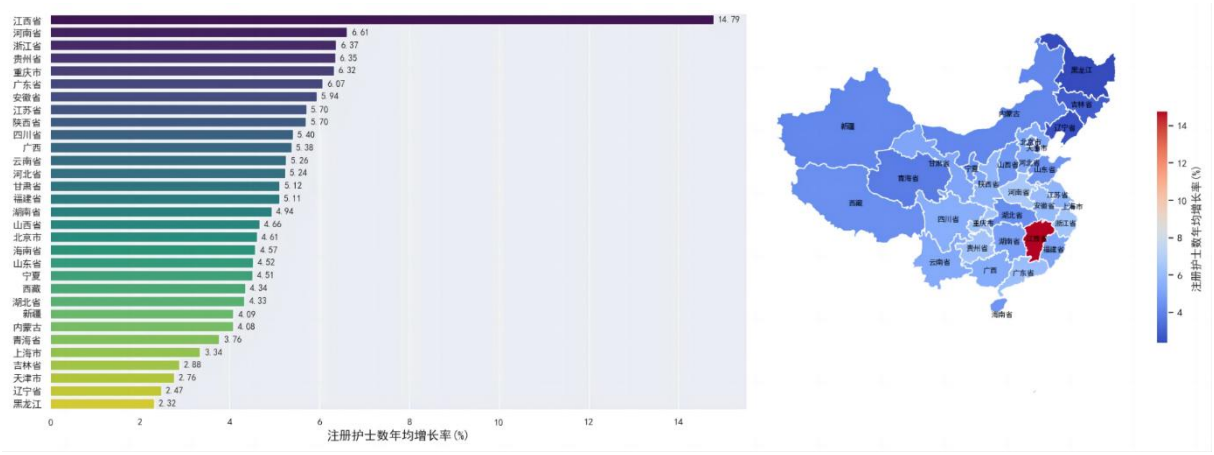


图 6-3：各地区注册护士数年均增长率可视化

图 6-3(左)展示了各省份注册护士数的年均增长率，图 6-3(右)平均增长率的热力地图反应其地理特征可以观察到几个关键点及其影响因素分析：

1. 江西省的年均增长率是第二大河南省的两倍多，浙江省、等省份年均增长率较大这可能反映出江西省在研究期间内对医疗卫生领域的投入加大，特别是在护士培养和招聘方面。
2. 黑龙江、辽宁省等省份年增长率较低，这些地区的增长率相对较低，可能受到人口流失、经济下滑以及医疗资源配置不均衡的影响。
3. 注册护士数增长较快的地区主要集中在东南地区，这些地区经济活跃，医疗卫生需求强烈，政策支持力度较大，从而促进了护士数量的快速增长。部分西北地区的增长率相对较低，西藏、青海等地，受地理位置和经济条件的限制，医疗卫生资源的增长相对缓慢。而东北地区最低。尤其是黑龙江和辽宁，可能由于经济下行压力、人口减少等原因，导致护士数量增长乏力。

6.3.4 卫生机构床位数差异及其可能影响因素分析

通过箱线图[5]可以观察到以下几个关键点及其影响因素分析：

1. 山东省、四川省与河南省等省份中位数较高。这些省份人口基数大，经济发展水平较高，医疗资源相对丰富，这可能是其床位数较高的主要原因。
2. 西藏、青海省与宁夏等省份中位数较低。这些地区地广人稀，经济发展水平相对较低，医疗资源相对匮乏，可能导致了床位数的不足。

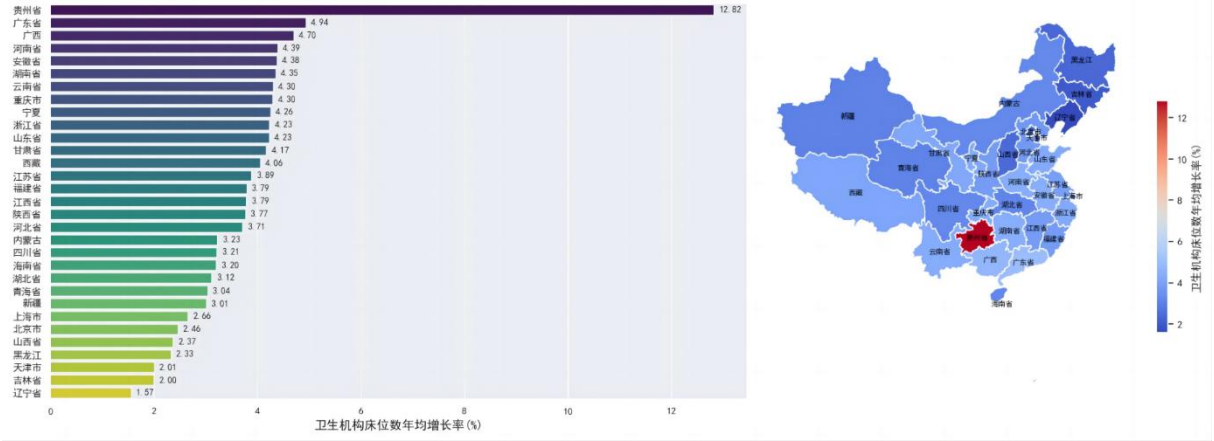


图 6-4：卫生机构床位数年均增长率可视化

[5] 卫生机构床位数箱线图见附录 1-4

图 6-4（左）展示了各省份卫生机构床位数的年均增长率，图 6-5（右）平均增长率的热力地图反应其地理特征可以观察到几个关键点及其影响因素分析：

1. 贵州省的年均增长率最高，几乎是广东省的三倍，广东省与广西的年均增长率也较高。这些地区的经济发展较快，人口流入量大，医疗需求旺盛，可能促使卫生机构加大投入扩充床位。
2. 辽宁省、吉林省与天津市的年均增长率相对较低，几乎在 2.00%以下。东北地区的经济增长相对缓慢，人口流失严重，可能影响了医疗资源的扩展速度。
3. 相对来说，北部地区增长率较低，东北尤甚。南部年均增长率较高，除海南省外。

6.3.5 医疗卫生机构数差异及其可能影响因素分析

通过箱线图[6]可以观察到以下几个关键点及其影响因素分析：

1. 四川省、安徽省等省份中位数较高，这可能与当地较大的人口基数、较高的经济发展水平以及政府的卫生投入政策有关。
2. 西藏、宁夏等较低，可能由于人口稀少、地理偏远以及经济发展相对滞后所致。

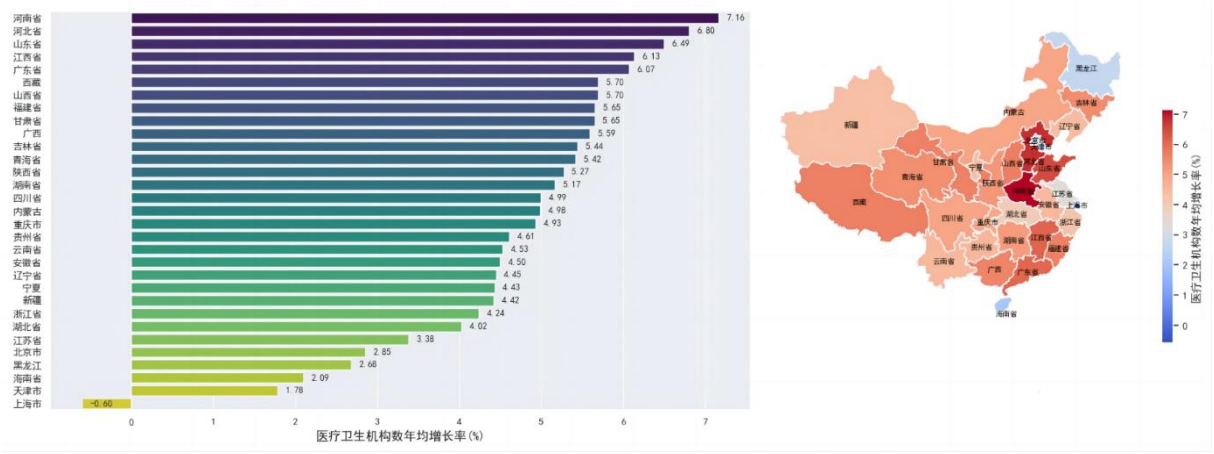


图 6-5：各地区医疗卫生机构数年均增长率可视化

图 6-5（左）展示了各省份医疗卫生机构数的年均增长率，图 6-5（右）平均增长率的热力地图反应其地理特征可以观察到几个关键点及其影响因素分析：

1. 河南省、河北省等医疗卫生机构数的年均增长率较高，上海、天津等较低，上海的年均增长率甚至呈负数-0.60%。这些城市的差异可能在于医疗卫生机构的饱和度。
2. 除了较北的黑龙江及较南的海南外，总体全国增长率较高，特别是中东部地区。这可能反映了经济发达地区和快速发展的中部地区对医疗服务需求的显著增长。

6.3.6 财政支出中卫生经费差异及其可能影响因素分析

通过箱线图[7]可以观察到以下几个关键点及其影响因素分析：

1. 广东省、北京市及江苏省等省份中位线较高，可能由于这些地区的人口密集、经济发达，政府有更大的财政能力投入更多的卫生经费，以提升公共卫生服务的质量。
2. 宁夏、海南省等则较低，可能是由于这些地区的财政能力较弱，经济发展水平相对滞后，导致公共卫生投入不足。

---

[6] 医疗卫生机构数箱线图见附录 1-5

[7] 财政支出中卫生经费箱线图见附录 1-6

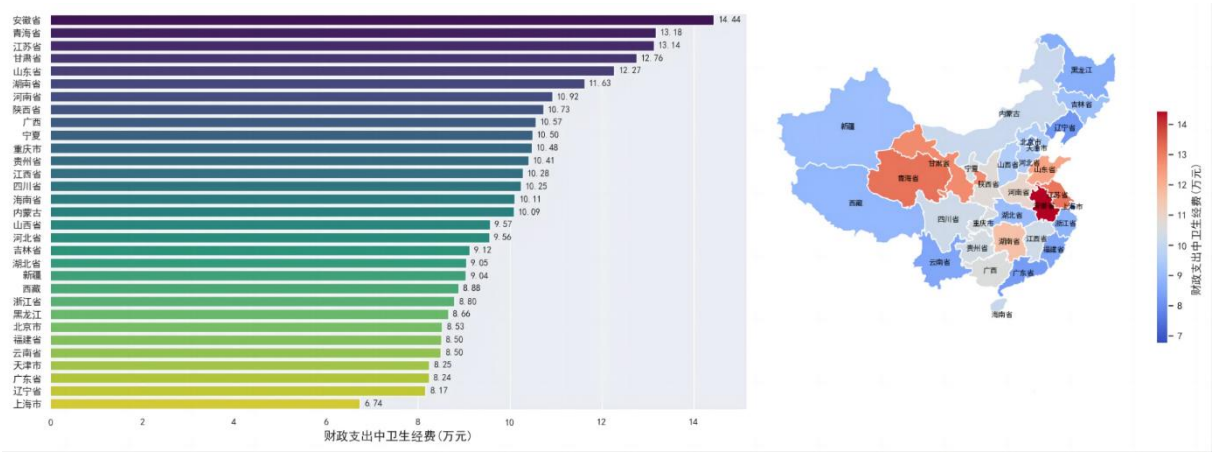


图 6-6：各地区财政支出中卫生经费年均增长率可视化

该图（左）展示了各省份财政支出中卫生经费的年均增长率，该图（右）平均增长率的热力地图反应其地理特征可以观察到几个关键点及其影响因素：

1. 安徽省、青海省及江苏省的财政支出中卫生经费年均增长率较高，而上海市、辽宁省及广东省较低，可能是因为这些地区已经具备相对完善的医疗卫生体系，财政支出的重点更多地放在优化和提升现有医疗资源的质量上。
2. 财政支出中卫生经费年均增长率高省份大致分布在中部地区。

6.4 综合评价

为了综合以上指标评价医疗资源配置，使用最大-最小标准化方法将数据进行标准化处理，计算各地区的综合得分。各指标比例分配如下表，因机构数为本背景主要研究对象，赋予更大比重：

表 6-2：指标权重分配

二级指标	三级指标	权重
医疗卫生人力资源	卫生技术人员数	0.1
	执业（助理）医师人员数	0.1
	注册护士数	0.1
医疗卫生物力资源	卫生机构床位数	0.2
	医疗卫生机构数	0.3
医疗卫生财力资源	财政支出中卫生经费	0.2

综合评价步骤如下：

1. 数据标准化。将数据合并为词典，见每个指标进行标准化。
2. 综合得分计算。赋予每个指标以上权重计算各省份每年的综合得分，使用热力图[8]展示每个地区的每年分数变化。

通过观察各地区 1990 年至 2023 年各年份的综合得分变化热力图（见附录）。总体来看各个地区的综合得分由低变高，全国的医疗资源配置呈逐步变好的趋势。

3. 计算每年得分并进行数据可视化，综合分数图表如下图 6-7：

[8] 见附录 3

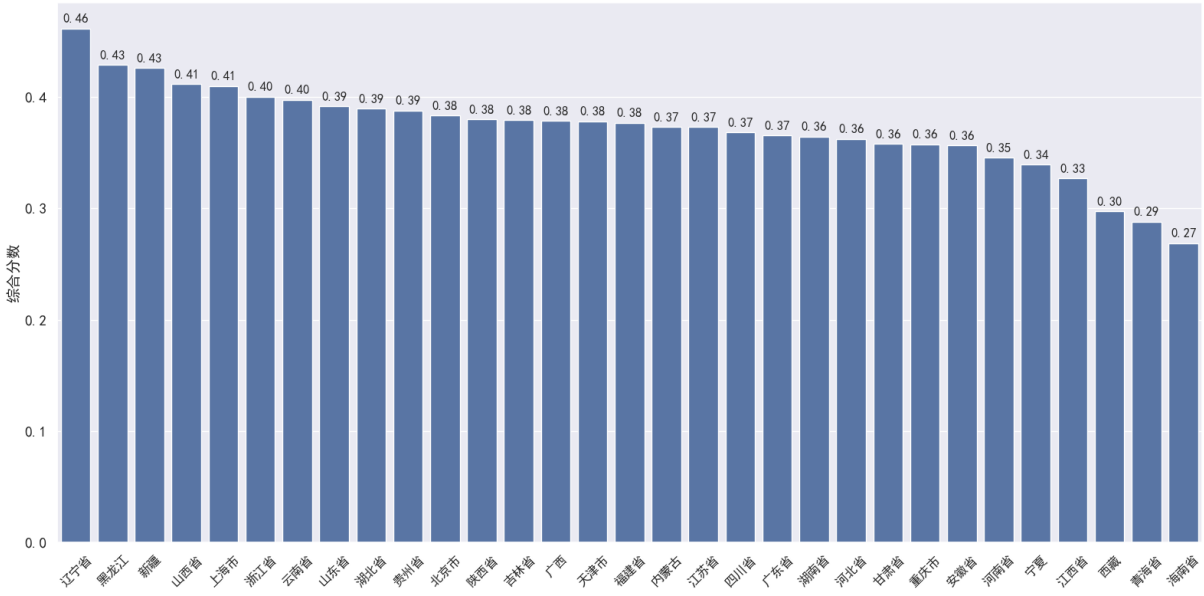


图 6-7：各地区医疗卫生资源配置综合得分

根据这个图表可以看出辽宁省、黑龙江等省份综合分数比较高，西藏、青海市与海南省的综合分数较低，但省份间指标差异较小，仅有 0.19。

根据以上结果综合评价得知，辽宁省、黑龙江省等得分较高的省份可能在人力资源配置方面有更强的优势，例如这些地区拥有更多的卫生技术人员、执业医师和注册护士。这些省份有较强的医疗教育资源，能够培养和吸引更多的医疗人才[7]。
西藏、青海和海南得分较低的省份可能因为地理位置偏远，医疗教育资源不足，导致人力资源配置的不足[8]。

辽宁和黑龙江等省份的高得分还可能与这些地区医疗卫生机构和床位数等物力配置的充分有关。这些地区的城市化进程相对较高，医疗设施建设相对完善，能够提供较为充足的物力资源支持。得分较低的省份则可能因为地广人稀，尤其是西藏和青海这类地广人稀的省份，医疗卫生机构数量和床位数难以达到较高的覆盖率，影响了整体资源配置得分。

财政投入高的省份通常在医疗卫生方面的财力支持较为充分，能够保障医疗设备、药品的供应，以及对医疗人员的激励和培训，这有助于提升医疗资源的综合得分。
得分较低的省份则可能因财政能力有限，难以在医疗卫生方面投入足够的经费，影响了医疗资源的可及性和服务水平。

6.5 背后的原因总结与政策意义

本文主要用六个医疗资源配置衡量指标数据将其可视化分析其可能影响因素，可以了解各省份医疗人力、物力及财力资源的配置情况与差异，分析其政策意义如下：总结主要有三种影响因素：

1. 经济发展水平。具有较高的经济基础的省份，如广东、江苏等，能够吸引更多的医疗卫生人员，且当地政府可以给予更多的卫生补贴、更充足的资源而经济相对落后的西北和东北地区在人力、物力、财力方面不足，医疗资源增长慢
2. 人口密度与需求。人口密集的地区，如广东省、山东省等，对医疗服务的需求高，使得政府增加医疗资源配置以满足需求。像人口稀少的地区，如西藏、青海，需求相对较低，医疗资源配置，如医疗卫生机构及医疗人员数不足。
3. 国家政策支持。国家对于中西部地区医疗资源配置的政策扶持起到了重要作用，如贵州的快速增长正是政策扶持的成果。

总结以上影响因素，对于了解如何再分配医疗资源、做政策决策尤为重要，具有政策意义。针对各省份的医疗资源配置差异，未来政策应该更加注重资源的再分配，尤其是针对经济发展较为落后的地区，提高全国的医疗资源配置的均衡度。人才引进与培养政策也尤为重要，在东北和西北等地区需通过政策优惠、生活条件改善等措施，增强对卫生专业人才的吸引力，避免人才流失，进而提升当地的医疗服务水平。针对一些经济发达地区，资源较为充足但增长率较低的情况，应该更多地关注资源质量的优化。

7. 问题三模型建立与求解

依据中国各地区医疗卫生机构数随时间变化折线图[9]，我们可以发现 31 个地区的变化曲线各不相同，但是一些地区在医疗卫生机构数的发展变化上呈现相似的模式；为了简化预测问题，我们依据数据变化的周期性、季节性和趋势，对各地区在 1990 年至 2023 年间医疗卫生机构数的变化模式做分类，共分成三类：曲线呈明显上升趋势且有小幅度波动、曲线趋势不明显且出现剧烈波动、曲线有上升趋势且发生较大波动。

针对各地区在医疗卫生机构数上的变化模式，我们选择三种不同模型进行预测。对于具有明显上升趋势且小幅度波动的地区，采用 Holt-Winters 模型进行机构数预测，其可以较好地对季节性成分做平滑处理及捕捉数据的增长趋势。对于变化趋势不明显且发生剧烈波动的地区，采用 SARIMA 模型进行预测，其能够有效地捕捉数据中季节性和非季节性特征，应对数据有规律或无规律的波动。对于有上升趋势且发生较大波动的地区，综合 Holt-Winters、SARIMA 和线性回归模型三者的预测结果进行最终预测，以期捕捉数据的上升趋势且平滑波动对预测产生的影响。

表 7-1：地区发展变化模式分类

发展变化模式	地区	模型
明显上升趋势且小幅度波动	北京，河北，山西，内蒙古，辽宁，吉林，黑龙江，江苏，浙江，安徽，江西，山东，河南，湖北，湖南，广西，重庆，四川，贵州，云南，西藏，陕西，新疆	Holt-Winters
变化趋势不明显且剧烈波动	天津，上海，海南，宁夏	SARIMA
有上升趋势且较大波动	福建，广东，甘肃，青海	Holt-Winters、SARIMA、线性回归组合模型

7.1 问题三模型建立

7.1.1 Min-max 标准化

为了降低 Python 程序运算复杂度，更好地提取数据本身特征，我们引入 Min-max 标准化方法；考虑在使用 Holt-Winters 模型时，可能会采用到乘法模型，需避免数据为 0 的情况，因此将数据按比例放缩到[0.1, 0.9]的范围内，公式如下：

[9] 详见附录 2

$$x' = 0.1 + \frac{(x - x_{min})(0.9 - 0.1)}{x_{max} - x_{min}} = 0.1 + \frac{(x - x_{min}) \times 0.8}{x_{max} - x_{min}} \quad (7-1)$$

x 为样本数据，在题境下为医疗卫生机构数，x_{min}为数据最小值，x_{max}为数据最大值，x'即为标准化值，范围在[0.1, 0.9]之间。

7.1.2 Holt-Winters 乘法季节性加法趋势模型

通过观察所选地区在过去 33 年医疗卫生机构数随时间变化的曲线，可以发现这一类地区都出现机构数量猛增的现象，且曲线中出现波动幅度各不相同，有的波动幅度逐渐增大，有的幅度逐渐变小，且在骤增阶段过后，数据保持平稳增长或降低的趋势。为了更好地拟合这种增长模式，我们选取 Holt-Winters 乘法季节性加法趋势模型，即在计算季节性成分上采用乘法模型，应对曲线中出现的波动幅度差异，计算趋势成分上采用加法模型，应对曲线末端平稳增长或降低的趋势，水平成分保持不变。

季节性成分的更新公式：

$$S_t = \gamma\left(\frac{y_t}{L_t}\right) + (1-\gamma)S_{t-m} \quad (7-2)$$

S_t是时间t的季节性成分；γ是季节性平滑系数，由程序根据数据自动优化得出，取值范围在 0 到 1 之间。y_t是在时间t上的实际观测值。S_{t-m}是t时间点对应的季节性因子，表示在季节周期m之前的季节性成分。

趋势成分的更新公式：

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \quad (7-3)$$

T_t是时间t的趋势成分；β是趋势平滑系数，由程序根据数据自动优化得出，取值范围在 0 到 1 之间。

水平成分的更新公式：

$$L_t = \alpha\left(\frac{y_t}{S_{t-m}}\right) + (1-\alpha)(L_{t-1} + T_{t-1}) \quad (7-4)$$

L_t是时间t的水平成分；α是水平平滑系数，由程序根据数据自动优化得出，取值范围在 0 到 1 之间。

综合三个成分的更新公式，预测公式为：

$$\widehat{y_{t+h}} = (L_t + hT_t) \times S_{t+h-m(k+1)} \quad (7-5)$$

$\widehat{y_{t+h}}$是时间$t+h$的预测值，h是预测的时间跨度，$k = \frac{h-1}{m}$表示预测时跨越了几个完整的季节性周期。

7.1.3 SARIMA 模型

为了挖掘存在巨大波动的数据的潜在模式，我们引入 SARIMA 模型对数据进行拟合并预测。SARIMA 由两部分组成，即非季节性部分和季节性部分。

非季节性部分的计算公式：

$$\phi(B)(1-B)^d y_t = \theta(B)\epsilon_t \quad (7-6)$$

该公式由三部分组成。在自回归部分，$\phi(B)$是自回归多项式，定义为$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$，其中$p$是自回归项的阶数，$B$是滞后算子。在差分部分，$(1-B)^d$是差分运算符，表示通过$d$阶差分使时间序列平稳化。在移动平均部分，$\theta(B)$是滑动平均多项式，定义为$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$，其中$q$是滑动平均项的阶数。$y_t$是在年份$t$时的数据，$\epsilon_t$是白噪声误差项。

季节性部分的计算公式：

$$\Phi(B^s)(1-B^s)^D y_t = \Theta(B^s)\epsilon_t \quad (7-7)$$

该公式由三部分组成。在季节性自回归部分，$\Phi(B^s)$是季节性自回归多项式，定义为$\Phi(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$，其中$P$是季节性自回归项的阶数，$s$是季节周期。在季节性差分部分，$(1-B^s)^D$是季节性差分运算符，表示通过$D$阶季节性差分去除季节性波动。在季节性移动平均部分，$\Theta(B^s)$是季节性滑动平均多项式，定义为$\Theta(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}$，其中$Q$是季节性滑动平均项的阶数。

结合非季节性和季节性两个部分，SARIMA 完整计算公式为：

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D y_t = \theta(B)\Theta(B^s)\epsilon_t \quad (7-8)$$

7.1.4 线性回归模型

为了挖掘时间序列数据中潜在的增长模式，我们引入了一次的线性回归模型，来优化拟合的结果。其计算公式为：

$$y = \beta_0 + \beta_1 X + \epsilon \quad (7-9)$$

β_0是截距，表示当$X=0$时y的值；β_1是斜率，表示X每增加一个单位时，y的变化量。ϵ是误差项，即无法解释的随机误差。

7.1.5 Holt-Winters、SARIMA、线性回归组合模型

我们将 Holt-Winters、SARIMA 和线性回归三个模型做组合，以期利用三者的特点拟合具有上升趋势且发生较大波动的数据。组合模型的公式如下：

$$\widehat{y_{mixed}}(t+h) = \frac{y_{Holt-Winters}(t+h) + \widehat{y_{SARIMA}}(t+h) + \widehat{y_{Linear}}(t+h)}{3} \quad (7-10)$$

代表将三个模型在数据集上的预测值取平均，充分结合每个模型自身的特点进行预测。

7.2 问题三模型验证

为了评估模型对所选地区数据的匹配性，反映模型的预测性能，我们选取$RMSE$、MAE和R^2三个指标来检验模型的拟合程度。

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (7-11)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (7-12)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (7-13)$$

n为样本数量，y_i为第i个样本的实际值，$\hat{y}_i$为第i个样本的预测值，$\bar{y}$为所有样本均值。

7.2.1 Holt-Winters 模型验证

为了检验 Holt-Winters 模型在所选地区数据上的匹配性，我们对数据集做分割处理，1990 年至 2020 年的医疗卫生机构数作为训练集，2021 年至 2023 年的机构数作为测试集，评估模型的预测效果；同时，为了提高模型的对未来预测的准确率，我们基于未来 5 年数据与过去 3 年数据误差分布相似的假设，选取MAE评价指标对预测数据做误差校正。最终，使用 2021 年至 2023 年的预测值和其实际值做拟合效果的评估来评价模型预测能力。评价结果如表 7-2 所示，拟合曲线如图 7-1 所示：

表 7-2：Holt-Winters 模型拟合效果

地区	$RMSE$	MAE	R^2
北京市	398.61	390.94	0.75
河北省	937.14	932.15	0.76
山西省	2621.20	2363.23	0.36
内蒙古	463.56	440.18	0.14
辽宁省	1521.62	1477.91	0.11
吉林省	691.93	629.61	0.13
黑龙江	201.40	193.77	0.66
江苏省	369.41	262.96	0.10
浙江省	761.25	663.91	0.14
安徽省	75.04	55.70	0.99
江西省	963.94	739.35	0.74
山东省	2003.20	1925.76	0.29
河南省	847.10	690.75	0.90
湖北省	112.85	101.01	0.98
湖南省	2394.13	2272.01	0.10
广西	861.16	819.05	0.31
重庆市	290.18	206.21	0.88
四川省	6279.37	5929.45	0.24
贵州省	566.34	565.47	0.36
云南省	179.06	154.84	0.95
西藏	150.10	135.15	0.23
陕西省	1232.07	1158.02	0.18
新疆	1539.54	1452.34	0.26

注：数值均四舍五入，保留两位小数

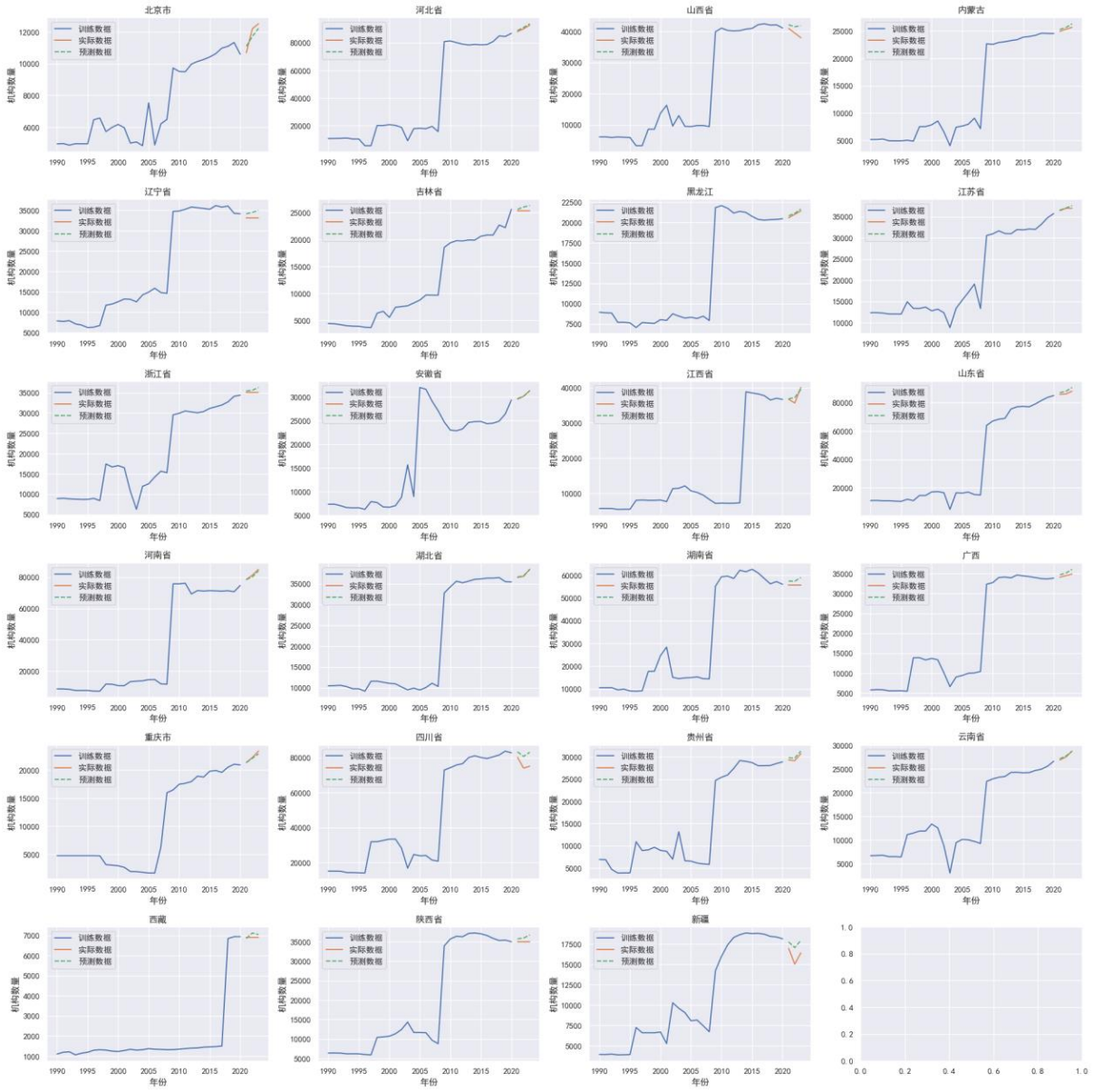


图 7-1：2021 年至 2023 年 Holt-Winters 预测值和实际值折线图

根据上述的评价结果可以看出，Holt-Winters 乘法季节性加法趋势模型能够较好地拟合预测所选地区的医疗卫生机构数，除部分地区存在偏差外，模型对大部分地区的预测数量基本吻合实际数据，且很好地预测所有地区的数据趋势，可以用作所选地区的预测模型进行未来 5 年医疗卫生机构数量预测。

7.2.2 SARIMA 模型验证

在问题一中，我们对数据集进行了线性填充，经观察，该部分所选地区 2022 年至 2023 年数据均缺失，为了提高预测的准确率，使用 2019 年至 2023 年共 5 年数据作为测试集，其他年份数据作为训练集，评估 SARIMA 模型的预测效果。在验证过程中，我们利用 python 的 pmdarima 库的 auto_arima 函数对每个地区数据搜索最佳的拟合参数，即确定最佳的非季节性自回归参数p、非季节性差分的阶数d和非季节性移动平均阶数q。SARIMA 在所选地区上的最佳参数如表 7-3 所示：

表 7-3：SARIMA 最佳参数

地区	p	d	q
天津市	0	1	1

上海市	2	0	0
海南省	0	0	0
宁夏	1	1	1

我们基于最佳参数设置的 SARIMA 模型，对每个地区进行预测分析，得到评价结果如表 7-4 所示，拟合曲线图如图 7-2 所示：

表 7-4：SARIMA 模型拟合效果

地区	$RMSE$	MAE	R^2
天津市	48.19	43.31	0.74
上海市	145.40	136.01	0.75
海南省	167.04	131.60	0.75
宁夏	137.11	130.16	0.17

注：数值均四舍五入，保留两位小数

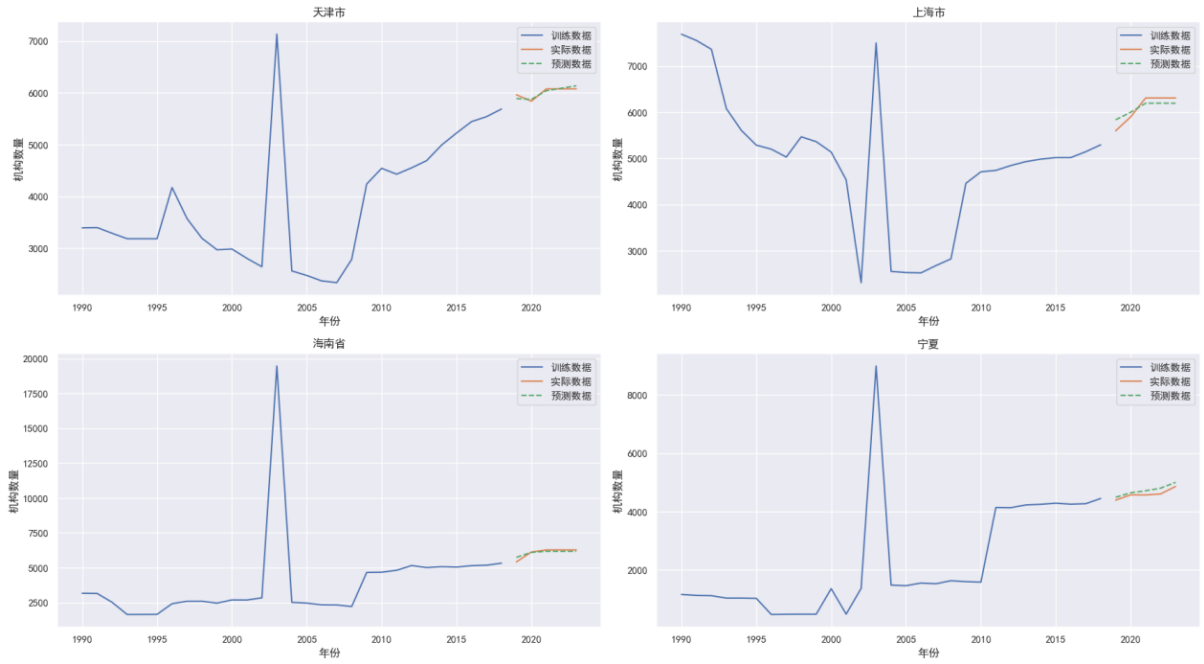


图 7-2：2019 年至 2023 年 SARIMA 预测值和实际值折线图

通过观察拟合效果数据及可视化结果，我们发现模型可以较好地预测所选四个地区未来的医疗卫生机构数，虽然模型对宁夏地区的R^2拟合度只有 0.17，但通过图表，我们观察到模型预测值基本符合实际数据的趋势，且能很好地消除历史巨大波动对数据预测产生的影响，SARIMA 模型可以用于该四个地区的医疗卫生机构数量预测。

7.2.3 组合模型验证

在数据集上，我们选取 2021 年至 2023 年共 3 年数据作为测试集，其他年份数据作为训练集。之后，我们采用 Holt-Winters、SARIMA 和线性回归构造组合模型，Holt-Winters 使用乘法季节性加法趋势模型，SARIMA 采用 auto_arima 函数进行最佳参数搜索来确定最优模型。SARIMA 在各地区数据上拟合的最优参数如表 7-5 所示：

表 7-5：SARIMA 最佳参数

地区	p	d	q
福建省	2	1	0
广东省	0	1	1
甘肃省	0	1	1

青海省	1	1	1

将三个模型的预测结果取平均，且采用MAE指标做误差校正，得到模型预测值与实际值的评估结果如表 7-6 所示，拟合曲线图如图 7-3 所示：

表 7-6：组合模型拟合效果

地区	$RMSE$	MAE	R^2
福建省	80.33	72.49	0.98
广东省	323.81	285.52	0.98
甘肃省	494.87	462.86	0.43
青海省	79.42	74.66	0.91

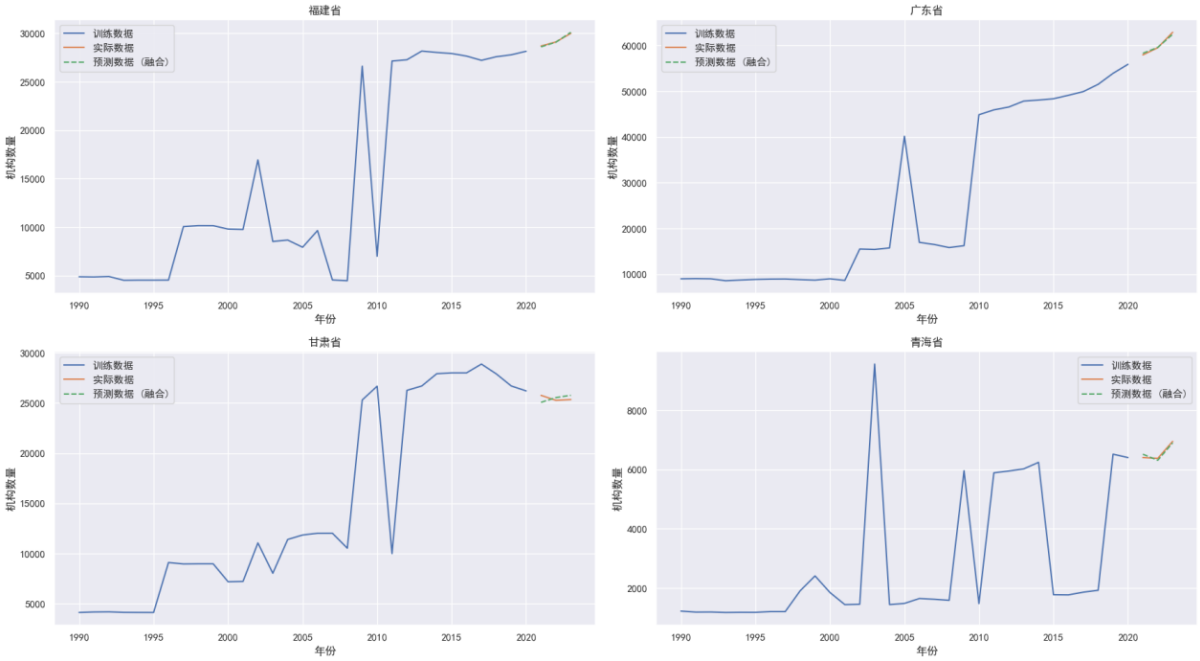


图 7-3：2021 年至 2023 年组合模型预测值和实际值折线图

通过观察拟合效果数据和可视化结果，可以发现组合模型很好地平滑了历史数据的巨大波动，并挖掘出数据潜在的增长趋势，预测值与实际值基本吻合。值得注意的是，虽然组合模型在甘肃省上R^2拟合度只有约 0.43，但和实际数据差异较小，可以用于甘肃省未来的医疗机构数预测。

7.3 问题三模型求解

求解过程均使用 Python 实现。首先，我们通过 scikit-learn 库中 MinMaxScalar 函数将数据放缩到[0.1,0.9]之间；然后基于建立的 Holt-Winters 乘法季节性加法趋势模型，我们使用 statsmodels 库中 ExponentialSmoothing 函数进行求解；基于建立的最佳参数 SARIMA 模型，我们使用 pmdarima 库中 auto_arima 函数进行求解；基于建立的组合模型，我们综合 ExponentialSmoothing 函数、auto_arima 函数和 scikit-learn 库中 LinearRegression 函数三者的结果进行求解，其中 auto_arima 函数使用在验证阶段得到的最佳参数组合。在获取到预测值后，我们使用在验证阶段统计的MAE指标对预测值进行校正，最终获取各地区在 2024 年至 2028 年的医疗卫生机构数如表 7-7 所示:

表 7-7：2024 年至 2028 年各地区医疗卫生机构数的预测值

年份 / 地区	2024	2025	2026	2027	2028
北京市	11944	12470	12393	12934	12842

河北省	94291	96904	99290	101909	104289
山西省	35936	37587	37854	39538	39772
内蒙古	25745	26602	26994	27863	28243
辽宁省	32309	33107	33840	34640	35372
吉林省	25625	26061	26913	27339	28201
黑龙江	21442	21991	22193	22750	22944
江苏省	37624	38600	39150	40136	40676
浙江省	34926	36107	36494	37693	38062
安徽省	31658	32753	33125	34238	34592
江西省	40323	41314	42411	43399	44499
山东省	87816	90816	92461	95494	97107
河南省	85027	89007	89601	93666	94175
湖北省	38798	40095	40483	41803	42168
湖南省	53605	56153	56295	58902	58984
广西	34571	35856	36321	37627	38071
重庆市	23835	24313	24968	25442	26101
四川省	69884	72862	73495	76532	77106
贵州省	30165	31600	31590	33059	33015
云南省	29114	29905	30447	31243	31779
西藏	7158	7132	7529	7491	7899
陕西省	34294	35557	36017	37301	37741
新疆	15331	15728	16083	16482	16836
天津市	6158	6249	6340	6431	6522
上海市	5983	5916	5815	5752	5688
海南省	4101	4101	4101	4101	4101
宁夏	4464	4464	4464	4464	4464
福建省	31253	32027	32822	33605	34392
广东省	63529	65297	67065	68833	70602
甘肃省	29419	30186	30954	31721	32488
青海省	5823	5921	6018	6115	6212

注：数值均四舍五入取整

7.4 结果讨论

根据模型评估的结果，模型预测值具备一定的可靠性，能够客观地反映数据的发展趋势，但仍然存在一些不确定性因素，这些不确定性主要由以下三个方面导致。

第一个方面就是数据质量与样本量，数据集中只有过去 33 年的历史数据点，一定程度上限制了模型的复杂度，导致模型在处理新数据时表现出较大的不确定性；且数据集中统计的历史 33 年数据本身会存在一定的统计误差，并不能真实地反映现实情况，这种系统误差也会导致模型预测结果不能很好地反映未来真实趋势。

数据质量的不足可能导致资源配置决策基于不准确的信息，从而影响政策的有效性。例如，如果某些地区的实际医疗需求被低估，可能导致这些地区的医疗资源不足，无法满足居民的健康需求。提高数据质量，通过增加数据样本量、改进数据收集和处理方法，减少统计误差，从而提高模型的预测精度。同时，在制定资源配置政策时，应考虑多个数据来源，并结合现场调查结果，确保决策的准确性。

第二个方面就是模型预测值校正方法的局限性，为了提高模型的预测能力，我们使用模型初次预测值与真实值的MAE指标作为校正误差，这一做法确实提高了模型的能力，但完全是基于模型对过去数据预测误差和对未来数据预测误差属于同分布的假设，一旦未来医疗卫生机构数的增长脱离模型拟合的增长模式，预测结果将出现极大的不确定性。

如果预测结果偏差较大，可能导致资源配置不合理，例如过度投资或资源浪费。为避免这种情况，政策制定者需要保留足够的灵活性，能够快速调整资源配置以应对意外情况。针对这个情况，除了依赖单一模型预测，还应结合多种预测模型，综合考虑不同的情景分析结果。同时，定期监测和评估实际数据，及时调整政策方向和资源配置方案，以应对变化。

第三个方面就是外部环境的变化；医疗卫生机构数的变化受到多种因素的影响，如人口密度、经济发展水平、政策变化等，这些因素的影响可能存在较大的不确定性，且模型本身采用的是单变量预测方法，仅从历史数据中总结出变化规律，而没有真正地考虑到外部因素，忽略了外部环境的动态变化，可能导致预测结果与实际情况出现偏差。

外部环境的变化可能导致某些地区的医疗需求急剧增加，而其他地区的需求则相对稳定或下降。如果模型未能准确捕捉这些变化，将可能导致资源配置不平衡，加剧地区间医疗资源的不均等。为了应对外部环境变化的不确定性，政策制定者应加强对外部环境的动态监测，尤其是对人口老龄化、经济增长和政策变化等关键因素的跟踪。同时，实施弹性政策，在资源配置上预留一定的调整空间，允许根据实际情况进行快速响应。

8. 问题四模型建立与求解

8.1 数据集扩充

医疗卫生机构数量受到众多外部因素的影响，参考吕烨在 2018 年对医疗机构需求的影响因素分析研究[15]，我们从四个方面考虑影响医疗卫生机构数的因素，即医疗卫生资源配置、人口结构、经济发展水平和政策。医疗卫生资源配置从卫生人员数和医药制造业专利申请数考虑，人口结构从总人口数、人口密度和 65 岁及以上人口比重考虑，经济发展水平从 GDP、人均 GDP、居民消费水平和城镇居民家庭人均医疗保健消费支出考虑，政策从政府卫生支出考虑。所选因素如表 8-1 所示：

表 8-1：影响医疗卫生机构数的因素

一级因素	二级因素	量纲
医疗卫生资源配置	卫生人员数	人
	医药制造业专利申请数	项
人口结构	总人口数	万人
	人口密度	人/平方公里
	65 岁及以上人口比重	%（百分比）
经济发展水平	GDP	亿元
	人均 GDP	元
	居民消费水平	元
	城镇居民家庭人均医疗保健消费支出	元
政策	政府卫生支出	亿元

针对所选取的因素，我们收集了这些因素从 1990 年至 2023 年的具体数据，数据缺失值数量如表 8-2 所示：

表 8-2：存在缺失值的因素及其缺失个数

存在缺失值的因素	缺失的年份个数
人口密度	1
卫生人员数	1
医药制造业专利申请数	6
居民消费水平	1
城镇居民家庭人均医疗保健消费支出	2
政府卫生支出	1

因缺失年份数量较少，参照问题一中数据缺失值填充方法，使用线性方法填充缺失值。

8.2 问题四模型建立

8.2.1 皮尔逊相关系数

为了识别所选因素和医疗卫生机构数之间的线性关系，同时可以排除无关变量，我们引入皮尔逊相关系数计算它们之间的相关性。计算公式如下：

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}} \tag{8-1}$$

r_{xy}是变量x和y之间的皮尔逊相关系数，取值范围为[-1, 1]。x_i和y_i为变量x和y的数据点，$\bar{x}$和$\bar{y}$分别为变量x和y的均值。

8.2.2 随机森林回归模型

为了评估所选因素对医疗卫生机构数的影响程度大小，我们引入随机森林回归模型来量化其影响程度。随机森林回归模型构建过程如下：

首先，对数据集的随机采样。对于训练集中的样本，随机森林使用有放回的抽样方法生成 100 个子样本集，每个子样本集用于训练一个决策树模型。

接着，对特征的随机选择。在构建每个决策树的过程中，随机选择部分特征用于寻找最佳分裂点。这种随机选择特征的方式可以降低模型的方差。

然后，训练决策树模型。对于每个随机采样的数据集，训练一个决策树模型。树的每个节点使用随机选择的特征来找到最佳分裂点。

最后进行预测。在进行预测时，随机森林集成所有决策树的预测结果，最终预测值是所有树预测值的平均值。

而随机森林中特征重要性的计算公式为：

$$\text{Importance}(X_j) = \sum_{t \in \text{Trees}} \sum_{s \in \text{Splits using } X_j} \Delta\text{MSE}(s) \tag{8-2}$$

$Importance(X_j)$表示特征X_j的重要性。$\Delta MSE(s)$表示在分裂点s处使用特征X_j所减少的均方误差，其定义如下：

$$\text{MSE}(t) = \frac{1}{N_t}\sum_{i \in t}(y_i - \bar{y_t})^2 \tag{8-3}$$

$$\Delta \mathrm{MSE} = \mathrm{MSE}(t) - \left(\frac{N_{t_L}}{N_t} \mathrm{MSE}(t_L) + \frac{N_{t_R}}{N_t} \mathrm{MSE}(t_R) \right) \quad (8-4)$$

N_t是节点t中样本的数量，y_i是样本i的目标值，$\overline{y_t}$是节点t中所有样本的目标值的均值。t_L和t_R分别代表左节点和右节点。

8.2.3 LSTM 模型

LSTM 模型是一种特殊的循环神经网络，其核心是它独特的单元结构，由输入门、遗忘门和输出门三个门控单元组成。

在每个时间步t，LSTM 接收输入数据x_t，上一个时间步的隐藏状态h_{t-1}，以及细胞状态c_{t-1}，然后生成新的隐藏状态h_t和新的细胞状态c_t，具体过程如下：

输入门控制当前输入x_t进行细胞状态的程度：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (8-5)$$

遗忘门控制上一个时间步的细胞状态c_{t-1}有多少被遗忘：

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (8-6)$$

细胞状态由遗忘门和输入门进行更新：

$$\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (8-7)$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t} \quad (8-8)$$

输出门控制细胞状态的哪一部分被输出为隐藏状态h_t:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (8-9)$$

最终隐状态h_t为：

$$h_t = o_t * \tan \mathrm{h}(C_t) \quad (8-10)$$

上述公式中，σ表示 Sigmoid 激活函数，输出值在(0, 1)之间；$tanh$表示 Tanh 激活函数，输出值在(-1, 1)之间；W_i, W_f, W_C, W_o是相应的权重矩阵；b_i, b_f, b_C, b_o是相应的偏置项。

单元细胞个数可以根据实际情况确定。我们在 LSTM 输出最后加入一个全连接层，控制模型输出值的维度：

$$y = W_{fc} \cdot h_t + b_{fc} \quad (8-11)$$

W_{fc}, b_{fc}分别为全连接层的权重矩阵和偏置项，y为模型的最终输出。

8.3 问题四模型求解

8.3.1 影响因素重要性分析

通过对每两个因素及因素与机构数间计算皮尔逊相关系数，得到因素相关性热力图如图 8-1 所示：

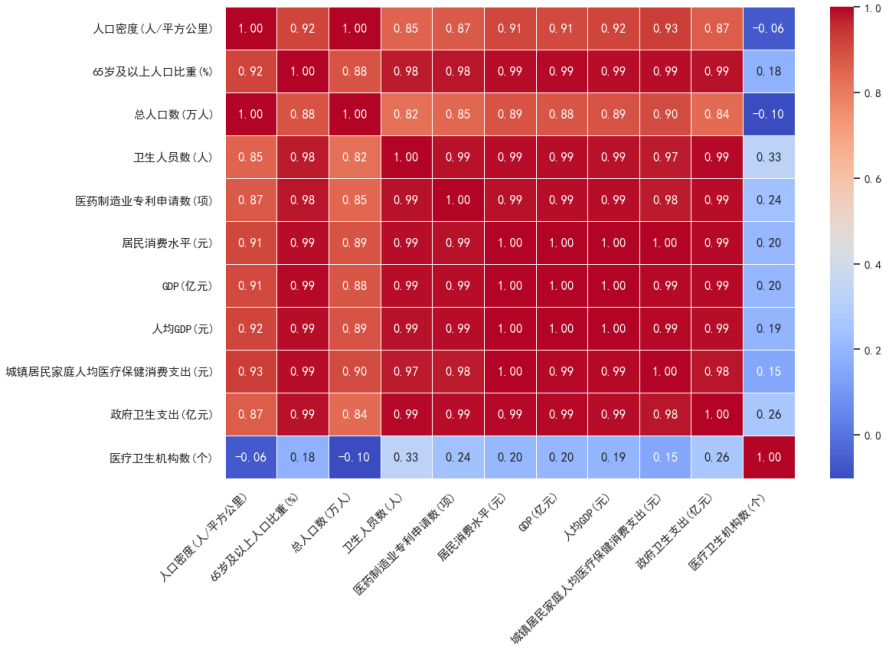


图 8-1：因素与医疗卫生机构数相关性热力图

在热力图中，红色越深代表正线性相关性越强，蓝色越深代表负线性相关性越强。从图中可以看出所选因素与医疗卫生机构数的线性相关性均较弱，结合已有信息可以推断出，中国医疗卫生机构数变化曲线呈凹形，即剧烈下降后骤增，而其他因素的变化均表现为逐步增长的趋势，因此单个因素与医疗卫生机构数的线性关系较弱。

其中，人口密度与医疗卫生机构数间的线性关系仅为-0.06，两者几乎没有任何的关系，我们将该因素做删除处理。值得注意的是，单个因素与目标量线性关系弱并不代表两者毫无联系，有可能两者存在非线性关系，也有可能是多个因素间的相互作用才对目标量产生影响，因此，我们继续采用剩下的 9 个因素对医疗卫生机构数做预测。

对随机森林回归模型求解，计算因素的重要性，得出因素重要性排序如图 8-2:

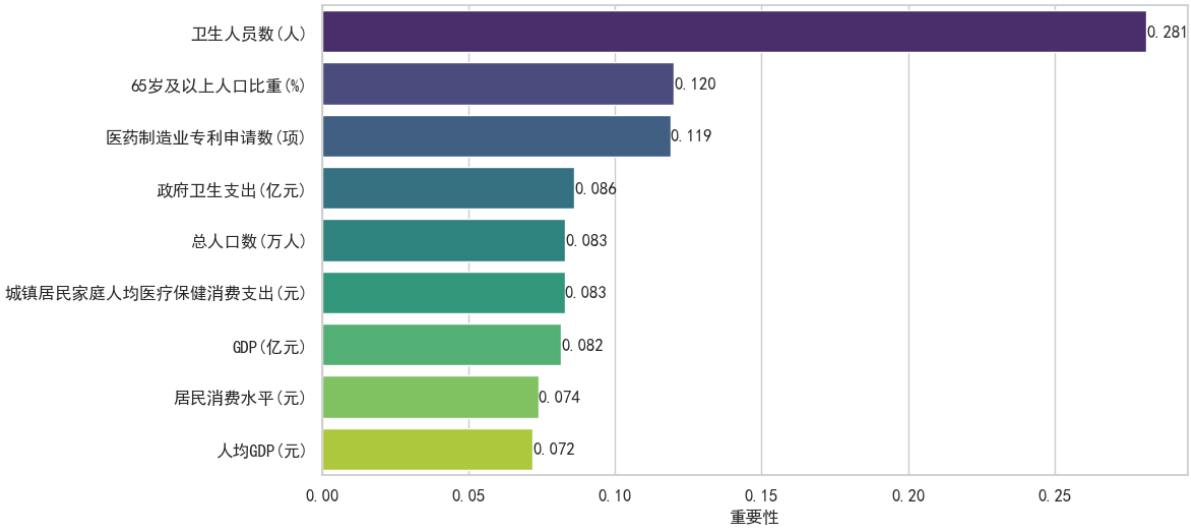


图 8-2：因素重要性程度

通过图 8-2 可以观察到，卫生人员数对医疗卫生机构数的影响最大，占约 28.1%的比重，其次是 65 岁及以上的人口比重和医药制造业专利申请数，相比之下，居民消费水平和人均 GDP 对医疗卫生机构数影响较小，分别为 7.4%和 7.2%。可以看出医疗卫生资源配置及人口老龄化问题对医疗卫生机构数影响显著，尤其是医疗卫生人员的配置，

很大程度上影响了医疗卫生机构的数量。

8.3.2 医疗卫生机构数宏观预测

8.3.2.1 LSTM 模型超参数调优

为了让 LSTM 模型能够更好地学习到数据的发展趋势，我们使用 Python 的 optuna 库对 LSTM 模型进行超参数调优，针对$lr, epochs, hidden, layers$，即学习率、训练轮数、隐藏层数和循环层数，搜索最佳的模型参数。将数据集按照 8：2 的比例划分训练集和测试集，优化目标为提高模型在测试集上预测值和真实值间的R^2指标。参数的调优范围如表 8-3 所示：

表 8-3：LSTM 模型超参数调优范围

超参数	最小值	最大值
lr	10^{-5}	10^{-2}
$epochs$	50	300
$hidden$	32	128
$layers$	1	3

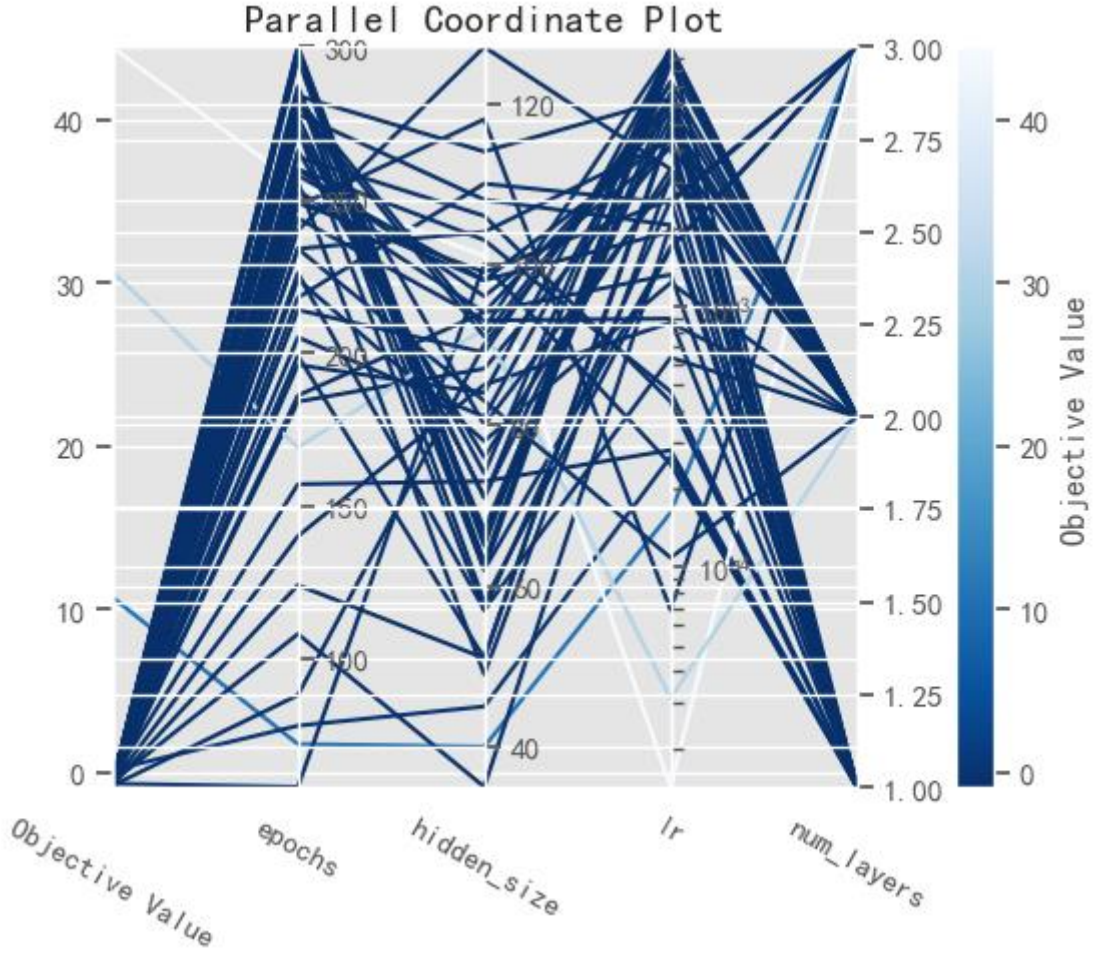


图 8-3：LSTM 模型超参数对结果的影响

经调优，最佳参数组合为$lr = 0.009596038092408676,\ epochs = 283, hidden = 92, layers = 1$。

8.3.2.2 LSTM 模型求解

基于最佳参数设置的 LSTM 模型，训练过程中损失函数下降曲线如图 8-4 所示：

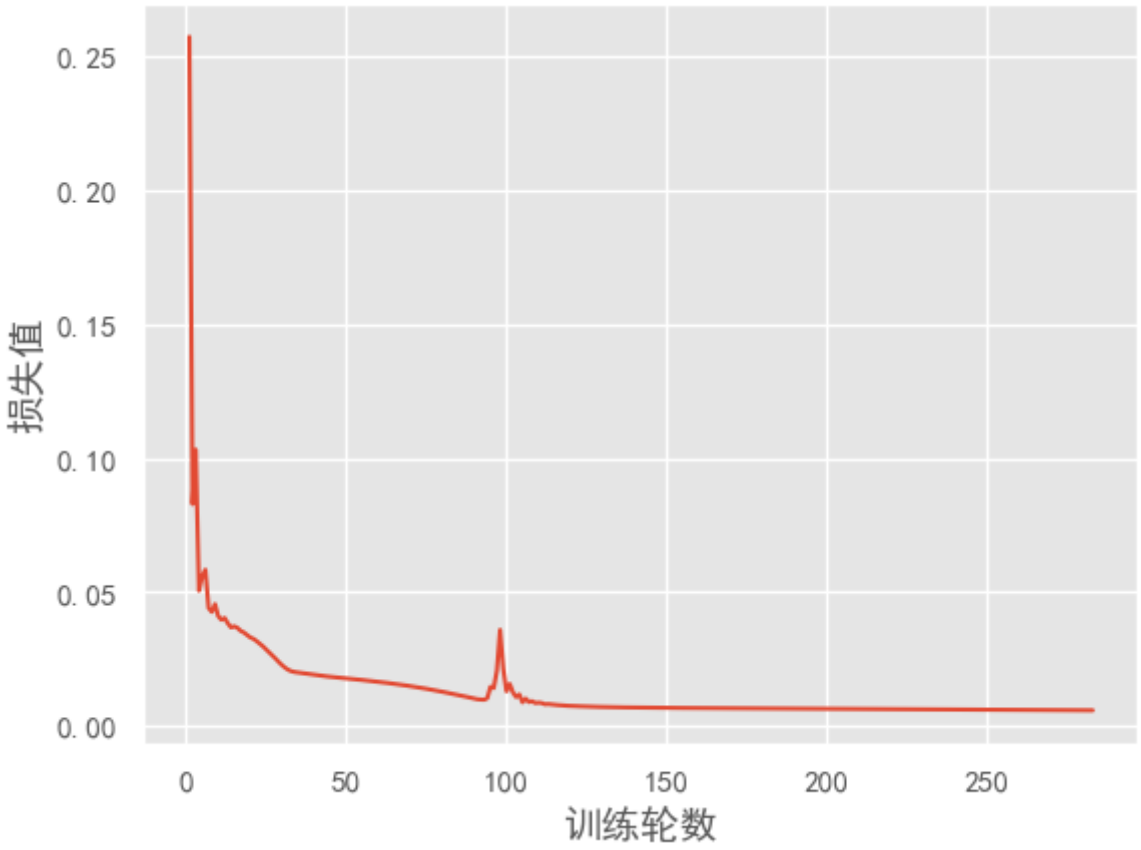


图 8-4：最优参数 LSTM 模型训练损失值曲线

可以看出，模型的损失值在前 50 轮训练中持续大幅度地下降，且在之后训练轮中保持缓慢地下降趋势，说明模型在训练过程中逐渐地被优化。将训练好的 LSTM 模型在测试集和未来 10 年的预测结果可视化如图 8-5 所示：

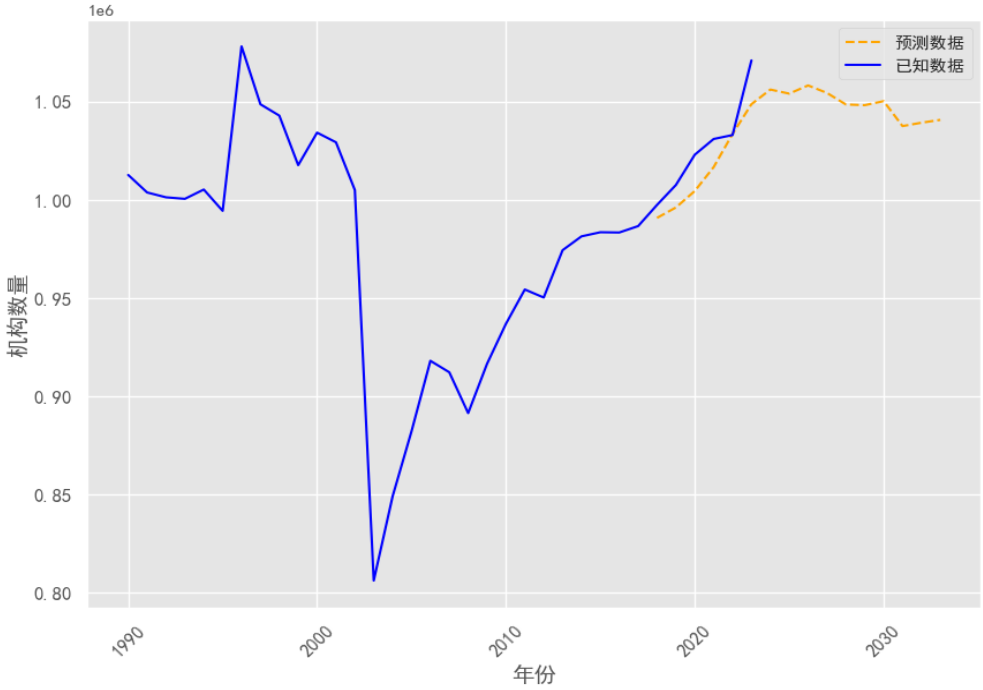


图 8-5：真实值和 LSTM 模型预测值曲线图

表 8-4：LSTM 模型对未来 10 年医疗卫生机构数预测值

年份	2024	2025	2026	2027	2028
医疗卫生机构数	1056155	1054077	1058250	1054426	1048565

表 8-4 续表：

2029	2030	2031	2032	2033
1048198	1050250	1037552	1039275	1040722

从图 8-5 中可以看出，LSTM 模型的对未来 10 年中国医疗卫生机构数预测值呈波动下降的趋势，但下降幅度较小。

8.4 结果分析

根据 LSTM 模型对未来 10 年医疗卫生机构数预测值可以看出，未来 10 年内，中国医疗卫生机构数量呈现波动趋势，且总体为下降趋势，期末略有上升。通过相关性分析可以看出医疗卫生资源配置及人口老龄化问题对医疗卫生机构数影响显著，尤其是医疗卫生人员的配置，很大程度上影响了医疗卫生机构的数量。

根据以上结论，我们得出以下对医疗卫生政策制定、资源分配和体系建设的启示：加强基层医疗卫生机构的建设。针对人口老龄化的趋势，政策走向应要加强在人口老龄化严重和经济较为落后的地区的医疗卫生机构的支持，如人力、物力与财力支持等，且鼓励优质医疗资源向相对落后的地区流动，促进医疗资源在不同地区的均衡发展[9]，以满足这些地区的需求。

推动医疗技术的创新与应用。根据以上结果可知，医药制造业专利申请数与医疗卫生机构也具有较强的关联性，所以继续推动新兴医疗模式的发展[10]，加强医院信息化建设、推动智慧技术研发等政策措施，将提供医疗机构技术支持，提高机构的服务质量，也有助于医疗卫生机构数量的增长。

促进优质资源扩容和区域均衡布局[11]。根据各地区的人口结构、经济水平的差别，制定对应的适合的医疗资源配置政策，确保资源的合理分布，避免过度集中或资源浪费。特别是对人口老龄化严重的地区，应加大医疗卫生机构的投入和建设，如建设国家区域医疗中心[12]，解决跨区域看大病、重病的需求，解决“看病难”的问题。

根据所得结果及其他资料，中国未来可能面临几个方面的挑战，以下为面临的挑战及其应对策略：

经济压力。在经济增长下滑的趋势下，政府对医疗机构的补助可能会下降，尽管医疗技术不断进步，但是慢性病率提高，高昂医疗费用将成为百姓的一个重大挑战。为了解决这个挑战政府可以鼓励社会资本参与[14]。

人才短缺。随着医疗机构数量的增长，及医疗服务需求的增加，对医疗机构专业人员需求也增加，医疗机构人员的短缺会降低医疗服务质量。针对这各问题，应通过提高教育和培训质量，吸引更多人才就业。

政策执行的不确定性。即使有良好的政策规划，但是在执行阶段会有落实协调困难、政策效果不佳的情况。针对这个情况，可以建立常态化、制度化沟通机制[15]，增加沟通透明度，提高沟通效率与准确度。强化监督制约，规范执行行为，将政策积极效果最大化。

参考文献

[1] 王俊豪,贾婉文.中国医疗卫生资源配置与利用效率分析[J].财贸经济,2021,42(02):20-35.DOI:10.19795/j.cnki.cn11-1166/f.20210207.003.

[2] 隆佩.人才吸引力影响因素的实证研究[D].上海财经大学:2022.DOI:10.27296/d.cnki.gshcu.2022.001549.

[3] 卫生机构，辽宁省人民政府办公厅关于印发辽宁省医疗卫生服务体系规划（2015—2020 年）的通知，https://wsjk.ln.gov.cn/wsjk/zfxxgk/fdzdgknr/lzyj/szfgfxwj/D04C2FB341EE4B5A887F00B318B3FE81/index.shtml，2024.8.23

[4] 张丰，贵州 5 项政策支持国家区域医疗中心建设，http://health.china.com.cn/2022-10/27/content_42150564.htm ，2004.8.23

[5] 国务院关于深入实施西部大开发战略情况的报告，http://www.npc.gov.cn/zgrdw/npc/zxbg/gwygyssxbdkfzlqkdbg/2013-10/22/content_1811909.htm，2024.8.24

[6] 浙江省发改委，省卫生健康委关于印发《浙江省医疗卫生服务体系暨医疗机构设置“十四五”规划》的通知，https://fzggw.zj.gov.cn/art/2021/6/29/art_1229123366_2306677.html，2024.8.24

[7] 黑龙江省人力资源和社会保障厅，人力资源和社会保障厅关于印发《黑龙江省加快数字人才培育支撑数字经济发展若干措施》的通知，https://www.echinagov.com/info/357158，2024.8.24

[8] 2024 两会笔谈，https://econ.pku.edu.cn/ztbd/lkbt/2024/376975.htm，2024.8.24

[9] 《中国智慧医院发展白皮书》，https://www.sohu.com/a/799254113_121993091，2024.8.24

[10] 央视网，国家卫健委等六部门：促进优质医疗资源扩容和区域均衡布局，https://news.cctv.com/2023/07/24/ARTIehTRtiQLSNybIwUW3uK8230724.shtml，2024.8.25

[11] 申少铁，推动优质医疗资源下沉（人民时评），http://opinion.people.com.cn/n1/2024/0614/c1003-40256075.html，2024.8.25

[12] 余央央.中国人口老龄化对医疗卫生支出的影响[D].复旦大学:2012.

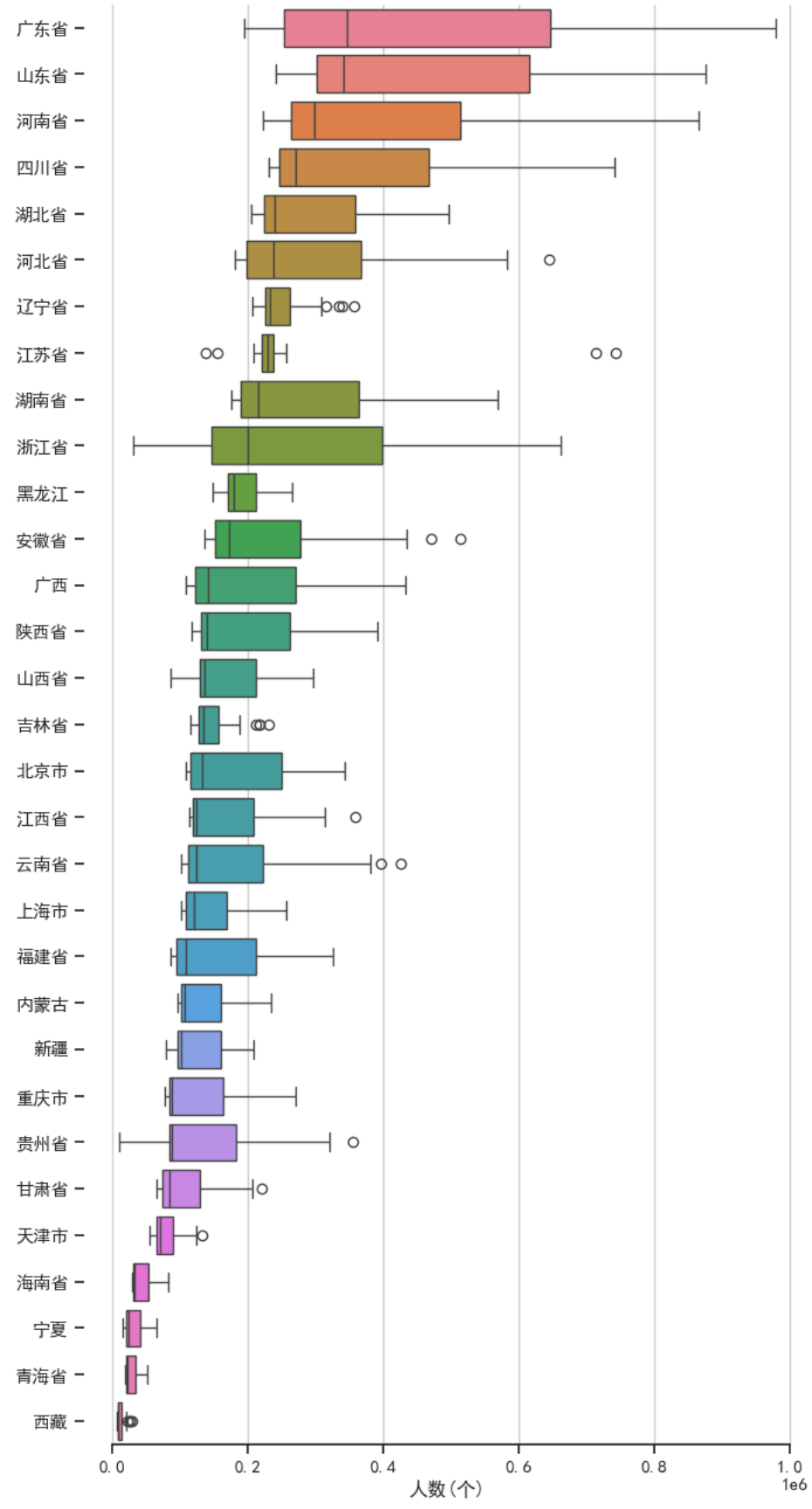
[13] 光明日报，医学科技创新与医学教育进步显著 为我国卫生健康事业发展提供有力保障，https://www.gov.cn/xinwen/2022-08/26/content_5706894.htm，2024.8.25

[14] 国务院，引导社会资本以 PPP 模式参与医疗机构建设运营，https://www.medsci.cn/article/show_article.do?id=90709214070，2024.8.25

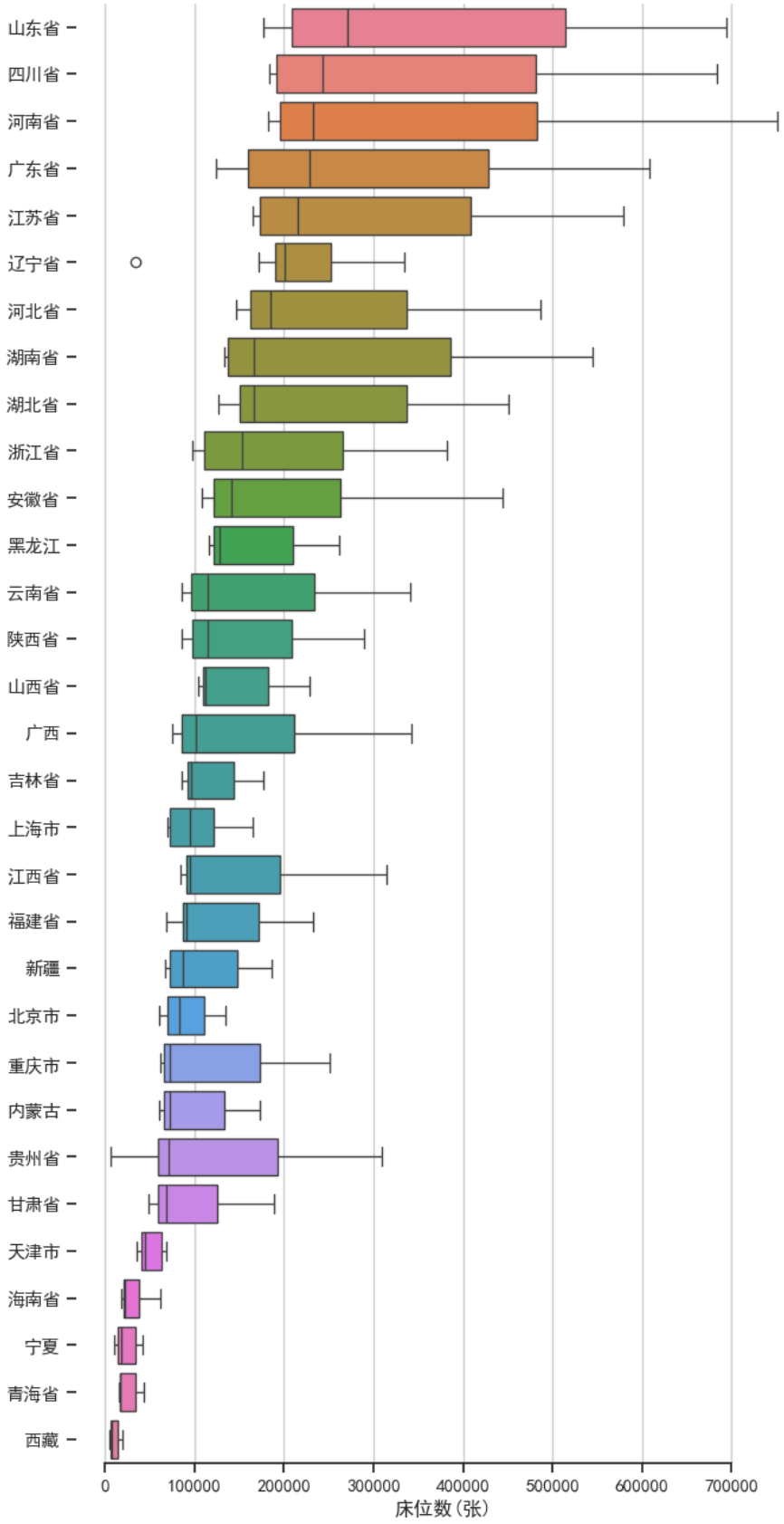
[15] 余新，精准培训提质增效，http://www.moe.gov.cn/jyb_xwfb/moe_2082/2021/2021_zl39/202105/t20210519_532248.html，2024.8.25

[15] 吕烨. 医疗机构需求的影响因素分析[J]. 统计学与应用，2018，7(1)：49-53.

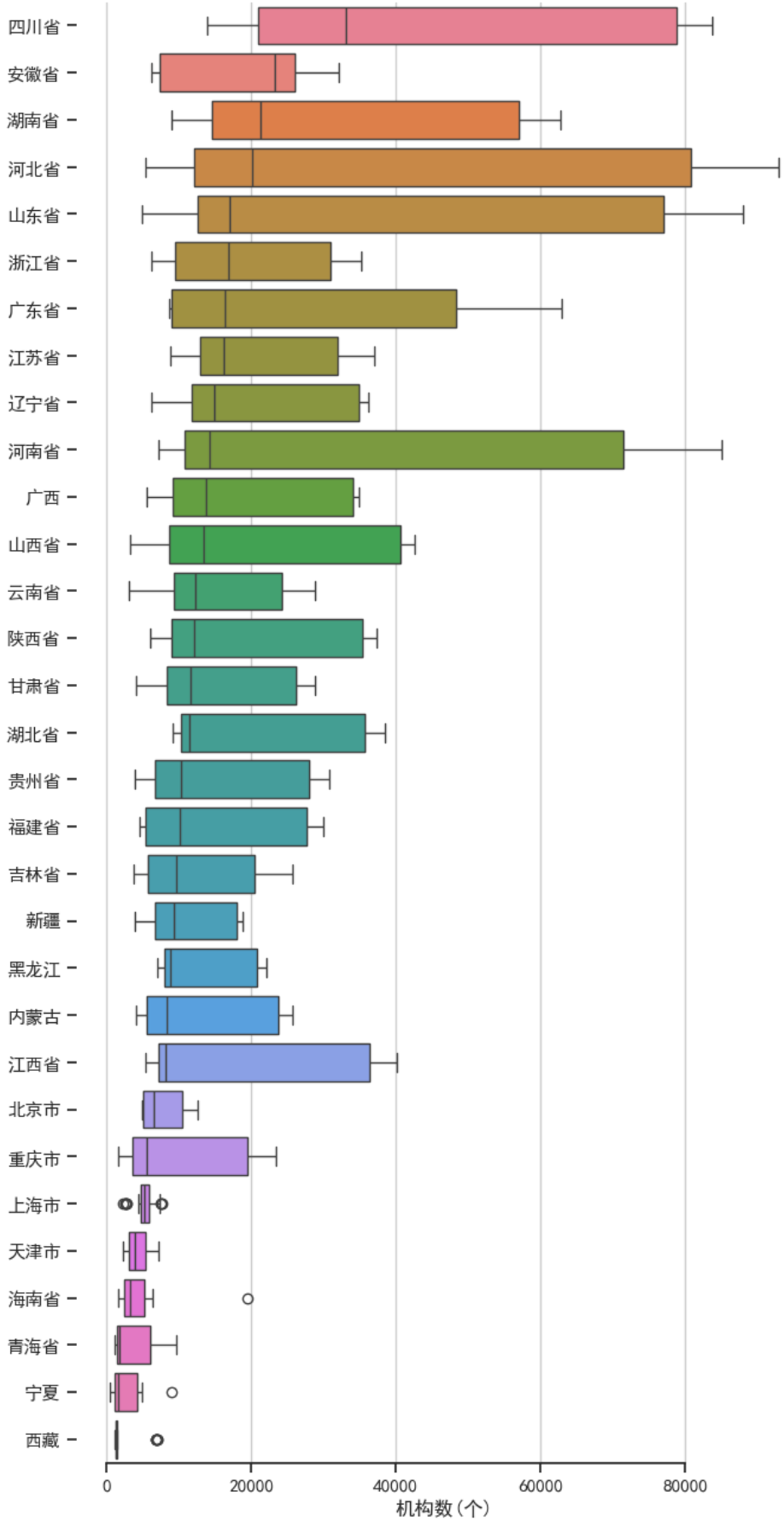
附录 1-1：


广东省
山东省
河南省
四川省
湖北省
河北省
辽宁省
江苏省
湖南省
浙江省
黑龙江
安徽省
广西
陕西省
山西省
吉林省
北京市
江西省
云南省
上海市
福建省
内蒙古
新疆
重庆市
贵州省
甘肃省
天津市
海南省
宁夏
青海省
西藏
0.0
0.2
0.4
0.6
0.8
1.0
1e6
人数（个）

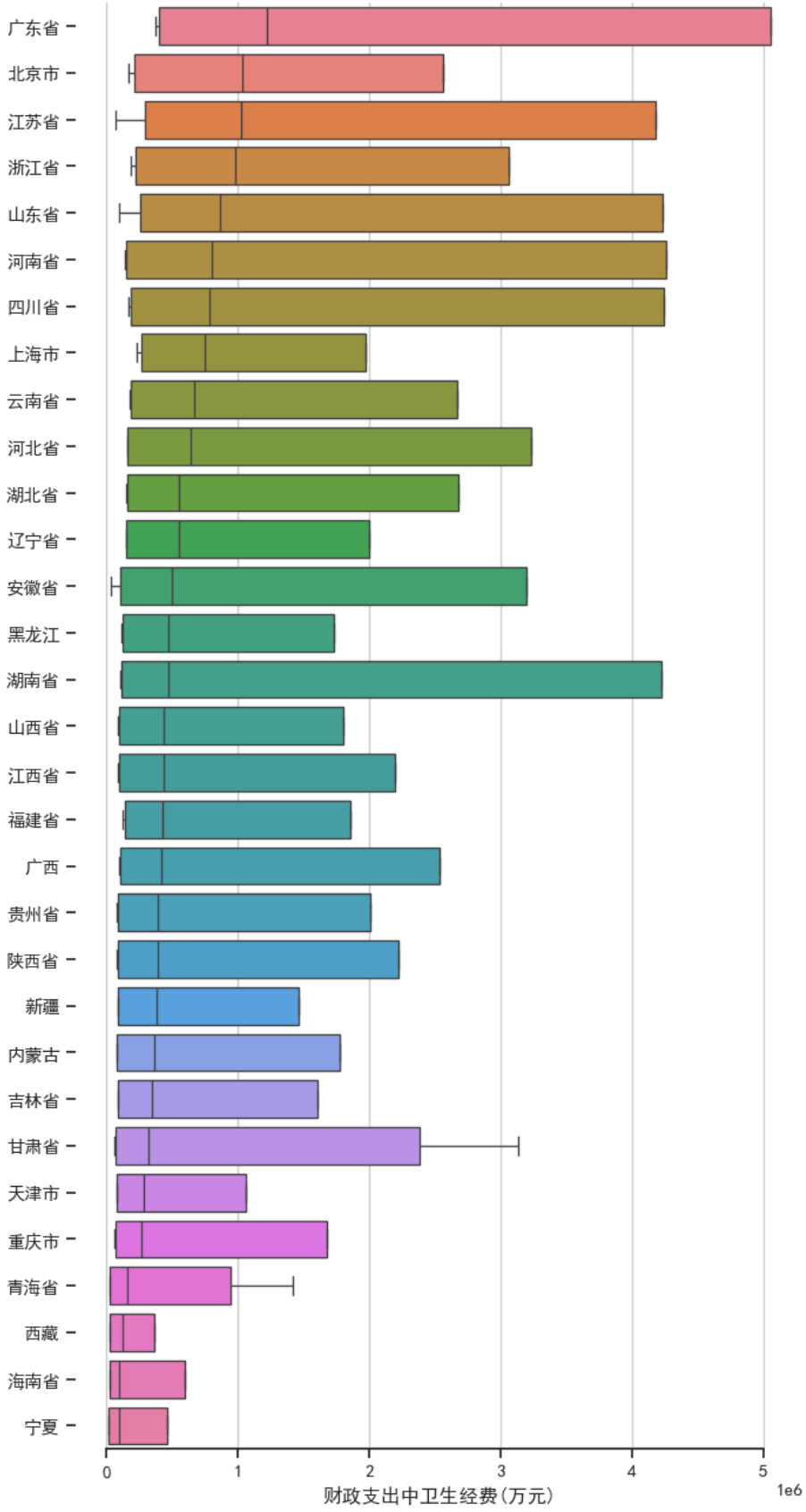
附录 1-2：


山东省
四川省
河南省
广东省
江苏省
辽宁省
河北省
湖南省
湖北省
浙江省
安徽省
黑龙江
云南省
陕西省
山西省
广西
吉林省
上海市
江西省
福建省
新疆
北京市
重庆市
内蒙古
贵州省
甘肃省
天津市
海南省
宁夏
青海省
西藏
0
100000
200000
300000
400000
500000
600000
700000
床位数（张）

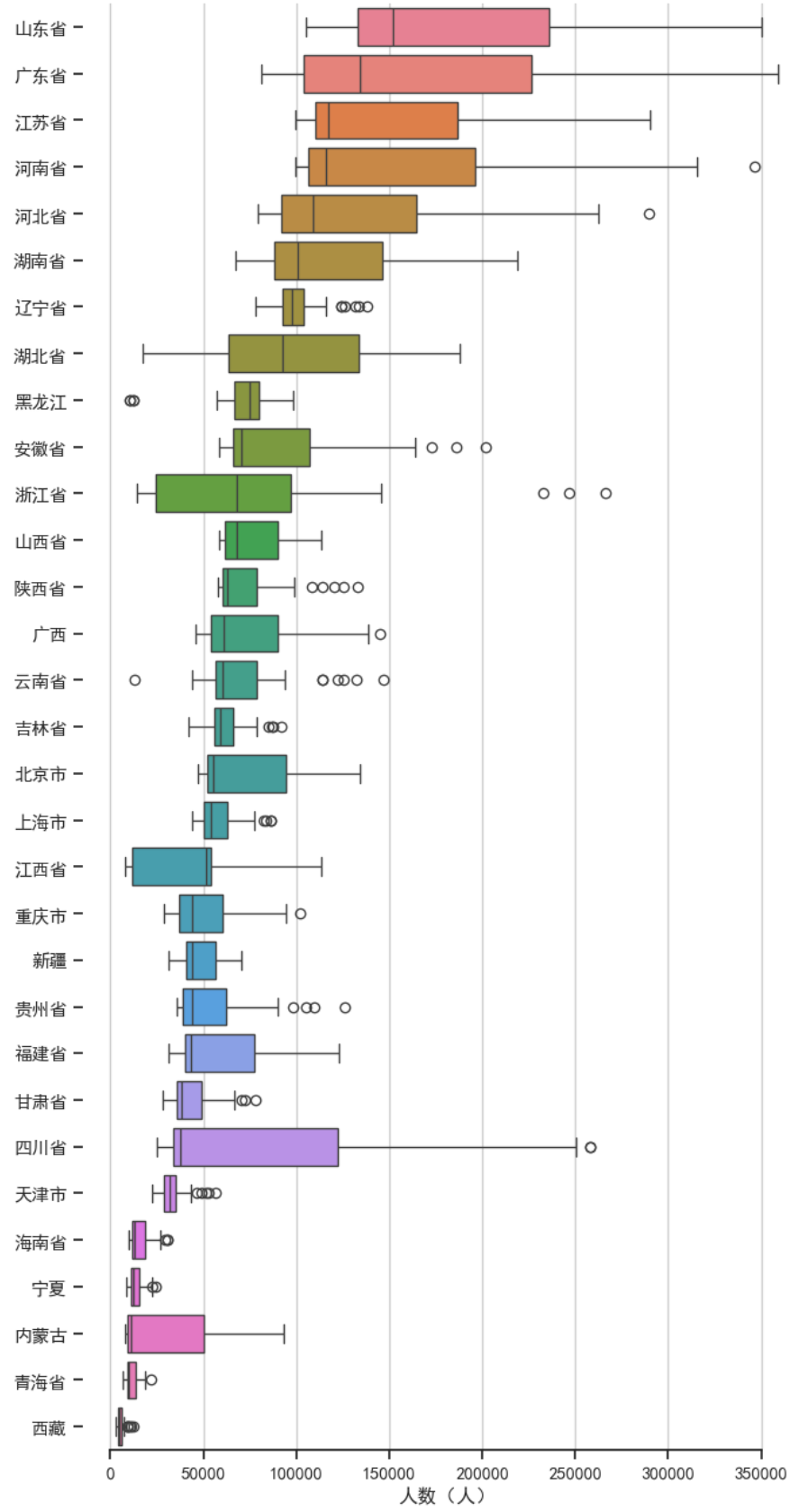
附录 1-3：


四川省
安徽省
湖南省
河北省
山东省
浙江省
广东省
江苏省
辽宁省
河南省
广西
山西省
云南省
陕西省
甘肃省
湖北省
贵州省
福建省
吉林省
新疆
黑龙江
内蒙古
江西省
北京市
重庆市
上海市
天津市
海南省
青海省
宁夏
西藏
0
20000
40000
60000
80000
机构数（个）

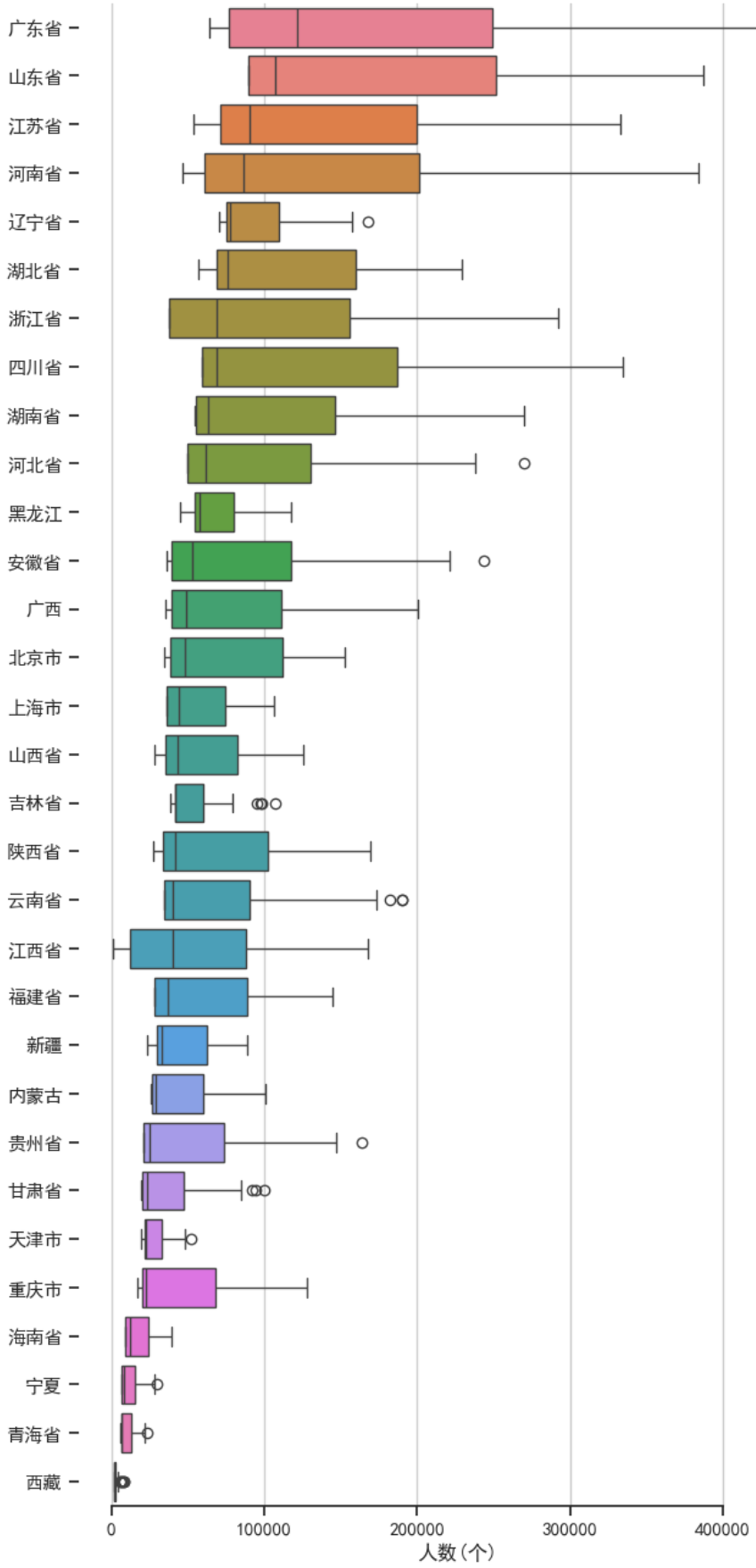
附录 1-4：


广东省
北京市
江苏省
浙江省
山东省
河南省
四川省
上海市
云南省
河北省
湖北省
辽宁省
安徽省
黑龙江
湖南省
山西省
江西省
福建省
广西
贵州省
陕西省
新疆
内蒙古
吉林省
甘肃省
天津市
重庆市
青海省
西藏
海南省
宁夏
0
1
2
3
4
5
1e6
财政支出中卫生经费(万元)

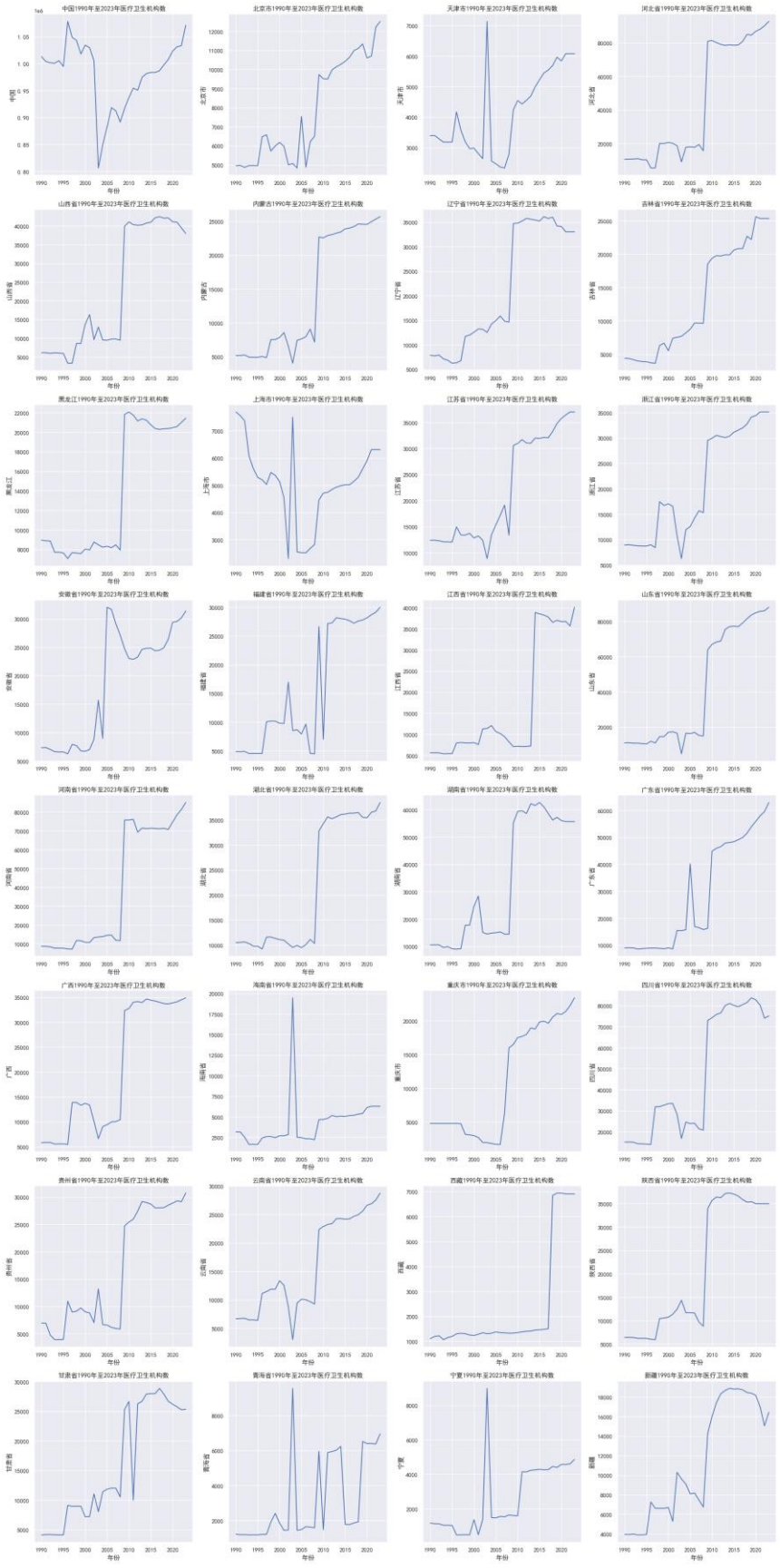
附录 1-5：


山东省
广东省
江苏省
河南省
河北省
湖南省
辽宁省
湖北省
黑龙江
安徽省
浙江省
山西省
陕西省
广西
云南省
吉林省
北京市
上海市
江西省
重庆市
新疆
贵州省
福建省
甘肃省
四川省
天津市
海南省
宁夏
内蒙古
青海省
西藏
0
50000
100000
150000
200000
250000
300000
350000
人数（人）

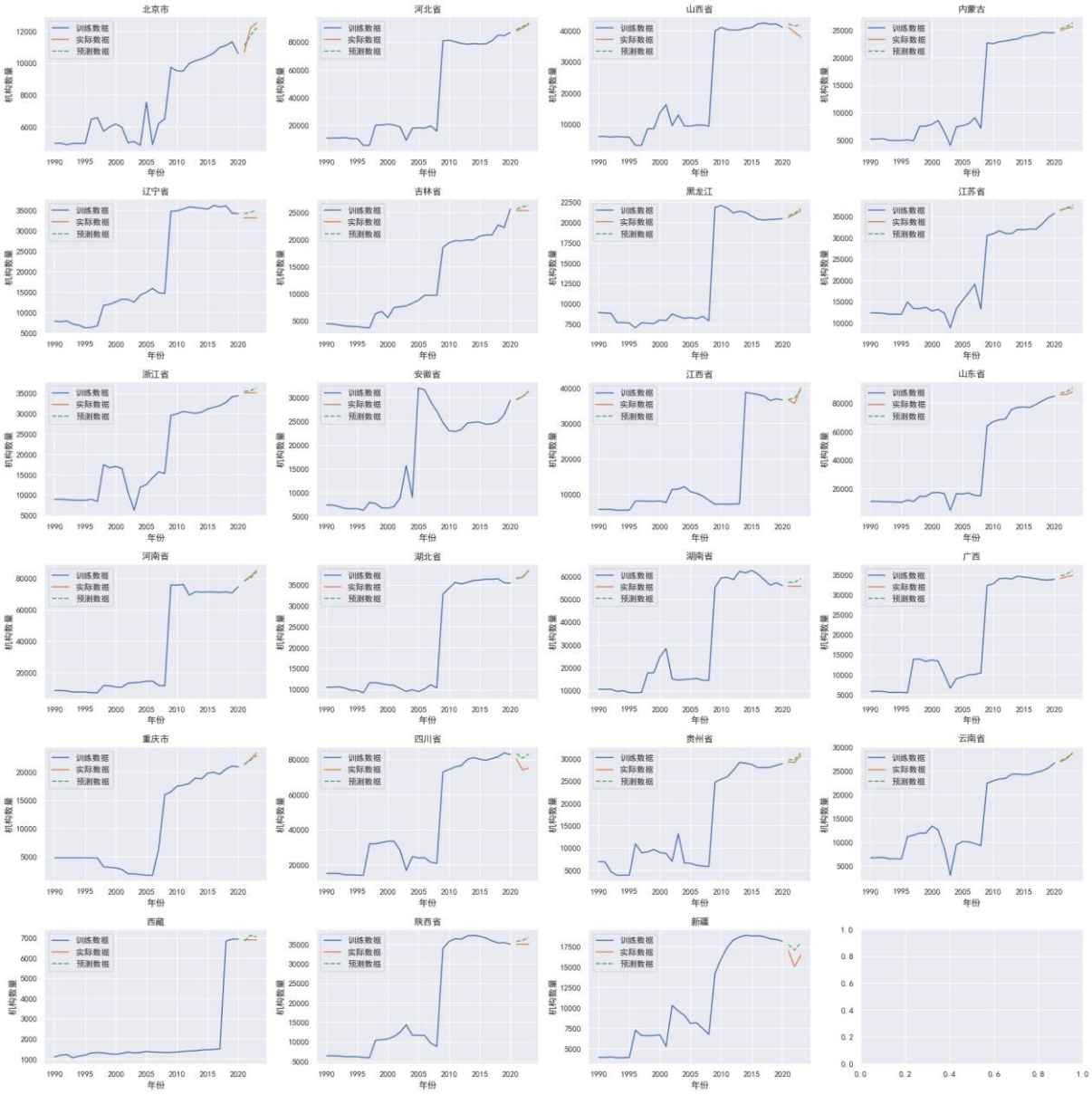
附录 1-6：

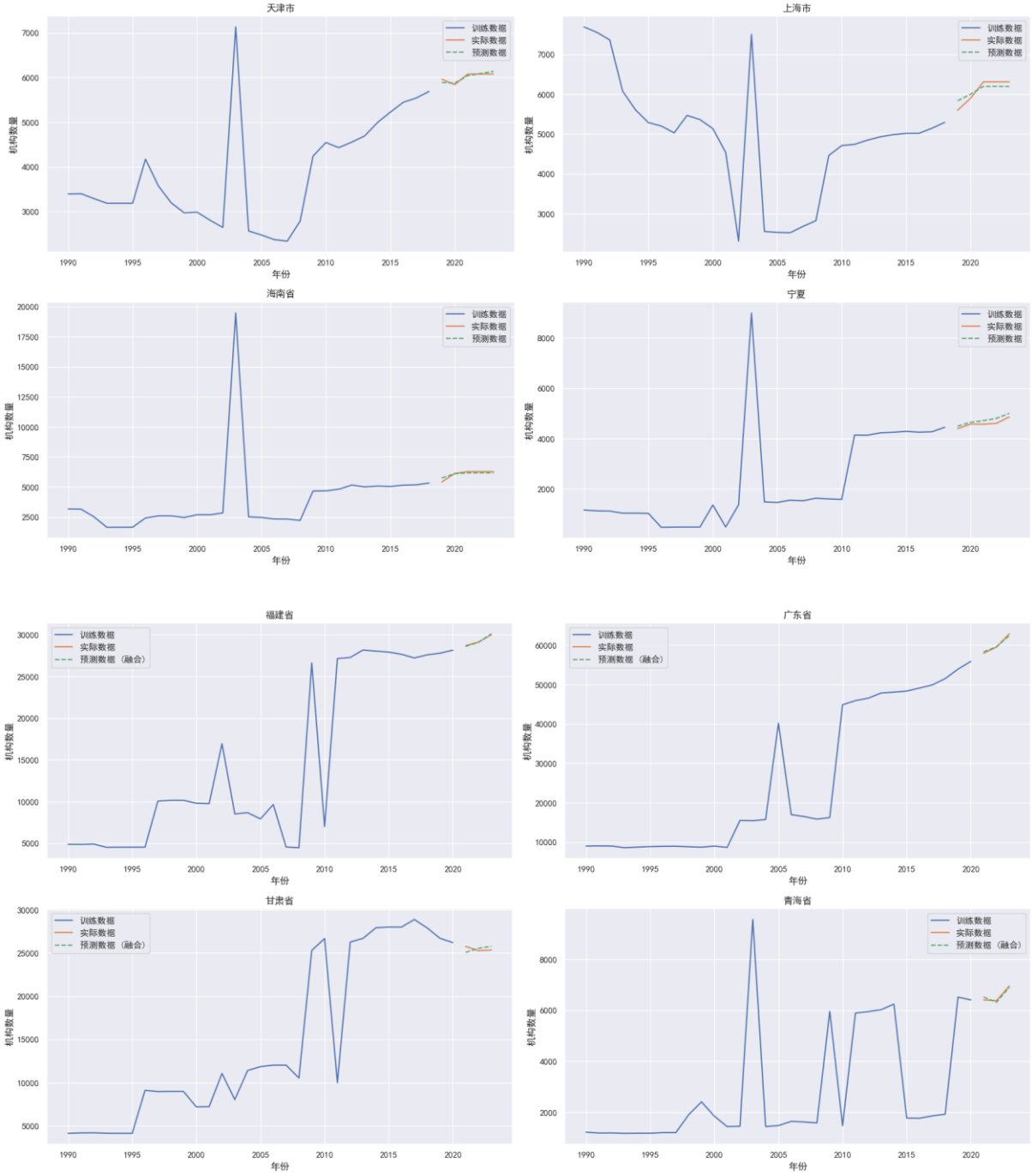

广东省
山东省
江苏省
河南省
辽宁省
湖北省
浙江省
四川省
湖南省
河北省
黑龙江
安徽省
广西
北京市
上海市
山西省
吉林省
陕西省
云南省
江西省
福建省
新疆
内蒙古
贵州省
甘肃省
天津市
重庆市
海南省
宁夏
青海省
西藏
0
100000
200000
300000
400000
人数（个）

附录 2：


中国1990年至2023年医疗卫生机构数
北京市1990年至2023年医疗卫生机构数
天津市1990年至2023年医疗卫生机构数
河北省1990年至2023年医疗卫生机构数
山西省1990年至2023年医疗卫生机构数
内蒙古1990年至2023年医疗卫生机构数
辽宁省1990年至2023年医疗卫生机构数
吉林省1990年至2023年医疗卫生机构数
黑龙江1990年至2023年医疗卫生机构数
上海市1990年至2023年医疗卫生机构数
江苏省1990年至2023年医疗卫生机构数
浙江省1990年至2023年医疗卫生机构数
安徽省1990年至2023年医疗卫生机构数
福建省1990年至2023年医疗卫生机构数
江西省1990年至2023年医疗卫生机构数
山东省1990年至2023年医疗卫生机构数
河南省1990年至2023年医疗卫生机构数
湖北省1990年至2023年医疗卫生机构数
湖南省1990年至2023年医疗卫生机构数
广东省1990年至2023年医疗卫生机构数
广西1990年至2023年医疗卫生机构数
海南省1990年至2023年医疗卫生机构数
重庆市1990年至2023年医疗卫生机构数
四川省1990年至2023年医疗卫生机构数
贵州省1990年至2023年医疗卫生机构数
云南省1990年至2023年医疗卫生机构数
西藏1990年至2023年医疗卫生机构数
陕西省1990年至2023年医疗卫生机构数
甘肃省1990年至2023年医疗卫生机构数
青海省1990年至2023年医疗卫生机构数
宁夏1990年至2023年医疗卫生机构数
新疆1990年至2023年医疗卫生机构数
年份

附录 3：模型对各地区预测值和真实值曲线图




天津市
上海市
海南省
宁夏
福建省
广东省
甘肃省
青海省
训练数据
实际数据
预测数据
预测数据（融合）
机构数量
年份

附录 4：各地区在 1990 年至 2023 年间医疗卫生资源配置综合得分