2019 年江西省研究生数学建模竞赛

参赛队号：　　20191038B

题目：某流行病致病原因分析与医疗保障政策建议

某流行病致病原因分析与医疗保障政策建议

摘　要

问题一要求选择合适的指标建立模型，分析该传染病在 2004-2016 年的流行病学变化趋势，并预测 2019 年全国感染该疾病的发病人数和死亡人数。考虑到题中强调的“2004-2016 年”这一信息，同时考虑到对于流行病学变化趋势和时间序列预测模型来讲，连续时间数据预测出的结果明显优于离散时间数据，故本问主要选取附件中的“总表”内数据进行分析。首先，对总表内数据如年份、发病数、死亡数等进行相关性分析，根据其结果判断可用于预测的数据。然后，考虑到总表内数据对于时间序列预测模型来讲并不属于高可靠性数据，故本问采用双模型并行预测——使用 BP 神经网络对每年总发病、死亡数进行预测；使用灰色马尔可夫模型对 A、B、C、D 四种方案诊断病例数进行预测。最终，分析对比两种模型的预测结果及预测误差，并选出最优值作为本问答案，即预测出 2019 年的发病数为 845134，死亡数为 2514。

问题二要求结合 2004 年,2007 年,2010 年,2013 年和 2016 年这五年按不同地区和职业分类统计的数据，建立该传染病模型，并且预测 2019 年传染病防控排名前 3 位的重点区域和重点人群。首先，针对这五年按不同地区和职业分类统计的该传染病的发病人数和死亡人数数据，应用地理信息软件 Arcgis10.4 进行全部和局部空间自相关分析，探测该传染病发病的时空聚集；然后依据这五年该传染病的时空分布特征，建立基于自回归求和移动平均（ARMIA）的该传染病传播模型；最后通过该模型，预测 2019 该传染病防控的区域和职业人群并进行排序，选择排名前 3 位的重点区域和重点人群。根据结果分析，得出重点区域分别为：西藏、新疆、青海；重点人群分别为：农民、家务及待业、离退人员。

问题三要求结合地区经济发展的相关公开数据，选择一个角度尝试建立该传染病与经济发展的数学模型，并分析得到结果。本问在问题一和问题二的基础上，通过查阅大量文献和资料，以广东省为例，查找了该省 2004、2007、2010、2013、2016 年共 5 年的人口结构、地区生产总值、城镇居民可支配收入、地方政府文教卫生支出、每 10 万人拥有卫生从业人员数等数据，分析数据相关性。然后，根据相关性结果选取出“流动人口”、“地方政府医疗支出”、“城镇化水平”、“居民人均可支配收入”四项指标，并提出 “可获得医疗帮助的能力”这一新的角度，以此建立多元线性回归模型。最后，使用该模型在广东省这 5 年内进行纵向对比，并与 2016 年上海传染病情况进行横向对比，以此来验证模型的可行性和可靠性。最终，通过与广东省数据对比，较好的体现了

本文模型的可行性和可靠性。之后又与其他省份的对比，模型的预测值依旧较为贴近实际值，体现了本文模型的普适性。

据题意可知，问题四需要结合前述问题分析结果，给卫生健康委员会相关部门写一封公开信，并给出对该传染病疫情防治的看法和建议。首先，对已完成问题和已有数据进行分析研判，由于农民、学生、离退人员、家政及待业等人群中发病数较多，且新疆、西藏、青海等地区发病率较高，故大致判断出该传染病应出现在抵抗力较弱、不注重平时保养、身体出现小问题不在意等群体身上，且较大可能出现在经济欠发达、地广人稀、医务人员稀少或流动人口过多的地区。其次，分析其他经济指标，发现有些政策以及政府地方财政等效果不明显，本问将主要针对这一情况作出相应分析和判断。最后，结合本文预测结果和分布趋势，给出对该传染病疫情防治的看法和建议。

关键词：BP 神经网络，灰色马尔可夫模型，ARIMA 模型，可获得医疗帮助能力

1 问题重述

1.1 问题背景

近年来，传染病监测预警已成为传染病防控的重点和热点研究问题，为了提高某传染病疫情和突发公共卫生事件报告的质量和时效，加强对全国感染病人的诊断、治疗和督导管理,卫生部建立了全国监管机制,及时通报相关病情和相关数据,并通过对疫情数据的动态分析，建立该传染病防治工作督导检查、防治效果评价和制定防治对策和策略，控制并逐渐消灭该传染病，提高全国人民的生活及健康水平。

传染病对全国人民的生活及健康的危害极大，随着信息技术的发展,以及传染病数据的大量增长,不少学者以大数据分析为基础对传染病进行监测预警。尤爱国等(2015)采用 Kruskal-Wallis 秩和检验比较不同类别传染病突发公共卫生事件持续时间及调查处置的投入时间，并分析得到 2013 年河南省学校传染病突发公共卫生事件主要是呼吸道传染病,其次是肠道传染病;春季和秋冬季为高发季节;农村小学是河南省学校传染病突发公共卫生事件发生的主要场所，以此提出综合措施进行防控[1]；祝丙华等(2016)基于网络、社会和自然因素、医疗、病原监测等不该传染病的措施来预防和控制同大数据来源的传染病监测预警系统对传染病进行监测[2]；曾子明等(2018) 建立基于 BP 神经网络的突发传染病舆情热度趋势预测模型预测突发传染病事件的发展趋势，进而对传染病以进行管控[3]；郭中凯等(2019) 研究了一类染病者具有年龄结构的 SIR 传染病模型的最优接种和治疗策略问题，借助切锥法锥技巧给出最优接种和治疗策略的必要条件[4]。通过监测数据对传染病进行分析与预测，对传染病的防治及提高全国人民的健康水平具有重要的现实意义。

1.2 问题要求

请通过对某传染病监管报告的部分数据(见原题附件 1)，主要内容包括年份、按诊断方法(A,B,C,D)、地区和职业分类统计的发病人数和死亡人数汇总数据，并结合必要的数据检索和扩充，对以下问题进行探讨。

- 问题一 分析该传染病病学变化趋势并预测 2019 年发病及死亡人数

选择合适的指标建立模型，分析该传染病在 2004-2016 年的流行病学变化趋势，并预测 2019 年全国感染该疾病的发病人数和死亡人数；

- 问题二 依据不同地区和职业建立模型并预测防控重点区域和人群

结合 2004 年，2007 年，2010 年，2013 年和 2016 年按不同地区和职业分类统计的数据，建立该传染病传播的数学模型，并预测 2019 年传染病防控排名前 3 位的重点区域和重点人群。

- 问题三 结合地区经济发展数据建立该传染病数学模型

结合地区经济发展的相关公开数据，选择一个角度尝试建立该传染病与经济

发展的数学模型，并分析你的结论；

- 问题四 给相关部门写一封公开信

结合上述讨论，给卫生健康委员会相关部门写一封公开信，谈谈您对该传染病疫情防治的看法和建议。

2 问题分析

2.1 问题一的分析

问题一要求选择合适的指标建立模型，分析该传染病在 2004-2016 年的流行病学变化趋势，并预测 2019 年全国感染该疾病的发病人数和死亡人数。考虑到题中强调的“2004-2016 年”这一信息，同时考虑到对于流行病学变化趋势和时间序列预测模型来讲，连续时间数据预测出的结果明显优于离散时间数据，故本问主要选取附件中的“总表”内数据进行分析。首先，对总表内数据如年份、发病数、死亡数等进行相关性分析，以此选择用于预测的数据。然后，考虑到总表内数据对于时间序列预测模型并不属于高可靠性数据，故本问采用双模型并行预测——使用 BP 神经网络对每年总发病、死亡数进行预测；使用灰色马尔可夫模型对 A、B、C、D 四种方案诊断病例数进行预测。最终，分析对比两种模型的预测结果及预测误差，并选出最优值作为本问答案。

2.2 问题二的分析

问题二要求结合 2004 年,2007 年,2010 年,2013 年和 2016 年这五年按不同地区和职业分类统计的数据，建立该传染病模型，并且预测 2019 年传染病防控排名前 3 位的重点区域和重点人群。首先，针对这五年按不同地区和职业分类统计的该传染病的发病人数和死亡人数数据，应用地理信息软件 Arcgis10.4 进行全局和局部空间自相关分析，探测该传染病发病的时空聚集；然后依据这五年该传染病的时空分布特征，建立基于自回归求和移动平均（ARMIA）的该传染病传播模型；最后通过该模型，预测 2019 该传染病防控的区域和职业人群并进行排序，选择排名前 3 位的重点区域和重点人群。

2.3 问题三的分析

问题三要求结合地区经济发展的相关公开数据，选择一个角度尝试建立该传染病与经济发展的数学模型，并分析得到结果。本问在问题一和问题二的基础上，通过查阅大量文献和资料，以上海市为例，查找了该市 2004、2007、2010、2013、2016 年共 5 年的人口结构、地区生产总值、城镇居民可支配收入、地方政府文教卫生支出、每 10 万人拥有卫生从业人员数等数据，分析数据相关性；然后，根据相关性结果选取出“流动人口”、“地方政府医疗支出”、“城镇化水平”、“居民人均可支配收入”四项指标，并提出 “可获得医疗帮助的能力”这一新的角度，以此建立多元线性回归模型。最后，使用该模型在上海市这 5 年内进行纵向对比，并与 2016 年广东省传染病情况进行横向对比，以此来验证模型的可行性和可靠性。

2.4 问题四的分析

据题意可知，问题四需要结合前述问题分析结果，给卫生健康委员会相关部门写一封公开信，并给出对该传染病疫情防治的看法和建议。首先，对已完成问

题和已有数据进行分析研判，由于农民、学生、离退人员、家政及待业等人群中发病数较多，且新疆、西藏、青海等地区发病率较高，故大致判断出该传染病应出现在抵抗力较弱、不注重平时保养、身体出现小问题不在意等群体身上，且较大可能出现在经济欠发达、地广人稀、医务人员稀少或流动人口过多的地区。其次，分析其他经济指标，发现有些政策以及政府地方财政等效果不明显，本问将主要针对这一情况作出相应分析和判断。最后，结合本文预测结果和分布趋势，给出对该传染病疫情防治的看法和建议。

3 符号说明及模型假设

3.1 符号说明

表 1 本文中出现的符号及其含义

符号	含义
$\bar{X}$	单个事件的均值
ρ	相关系数
S_j	隐藏层各神经元的激活值
w_{ij}	输入层至隐藏层连接的权值
θ_j	隐藏层单元的阈值
$\Delta^0(t)$	相对残差
N	各省农村人口
C	各省城镇人口
$N+C$	常住人口
Y	城镇化率
S	人均收入
SC	城镇人均收入
SN	农村人均收入
S	居民人均可支配收入

3.2 模型基本假设

1.题中所给内容和数据资料都是真实可信的；

2.国家统计年鉴数据真实可靠；

3.地方政府统计年鉴数据真实可靠；

4.每三年的统计数据也可视为连续时间序列数据；

4 模型的建立与求解

4.1 数据预处理

4.1.1 异常值检测与处理

做出随年份变化的趋势图（地区、职业、分诊断方案的），如图 1 所示为 2004-2016 年该传染病发病、死亡数示意图（其余类似趋势图见附录）。

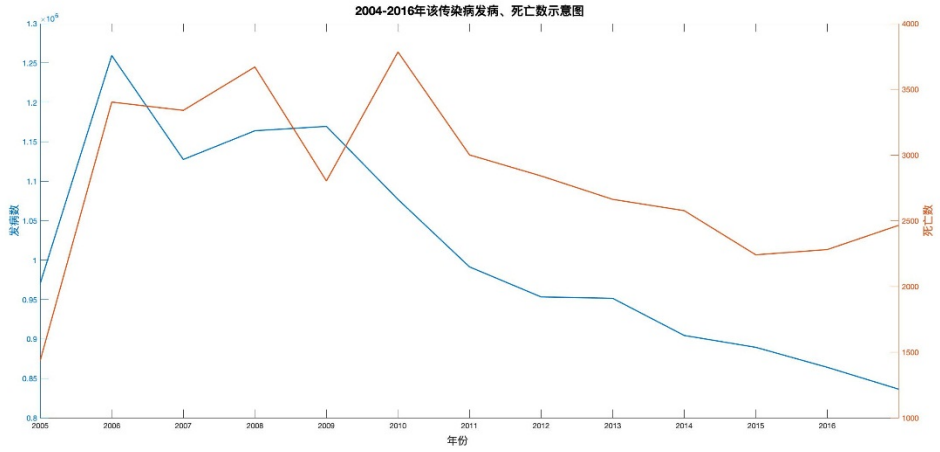

图 1　2004-2016 年该传染病发病、死亡数示意图

观察其走势，未发现异常数据点，故判断无异常值。不进行异常值处理。

4.1.2 空白值检测与处理

发现空白值（如附件 1 内 2014 年甲中）西藏、宁夏在诊断方法 A 时其发病人数和死亡人数为空白，如表 1 所示。

表 1 原题所给数据的空白值举例

地区	诊断方法 A		诊断方法 B		诊断方法 C		诊断方法 D	
	发病数	死亡数	发病数	死亡数	发病数	死亡数	发病数	死亡数
西藏			137	0	821	2	983	1
陕西省	21	0	15148	30	8732	16	9558	15
甘肃省	18	0	6195	14	9107	6	4613	4
青海省	2	0	1672	1	991	2	1208	1
宁夏			1124	0	1668	0	1067	0
新疆	32	0	9397	15	8617	15	8375	2

由于年份数据过短、数据过少，故部分缺失数据无法用插值法补全。同时，由于缺乏年份大多较早，很可能存在该种诊断手段尚未推广至该地区导致数据空缺的可能，且数值极小，难以对模型结果产生影响，故缺失部分直接舍去。

4.1.3 反归一化值

（1）由于问题一中需要使用到具体数字，故不采用反归一化；

（2）从问题二开始，由于传染病学中习惯使用“/10 万人”为量纲，故本文自第 2 问开始，采用反归一化值，单位为“/10 万人”。某年内某地发病数的反归一化值=$\frac{\text{该地发病数(人)}}{\text{该地人口数(万人)}}\times 10$

4.2 问题一模型的建立与求解

问题一要求选择合适的指标建立模型，分析该传染病在 2004-2016 年的流行病学变化趋势，并预测 2019 年全国感染该疾病的发病人数和死亡人数。考虑到题中强调的“2004-2016 年”这一信息，同时考虑到对于流行病学变化趋势和时间序列预测模型来讲，连续时间数据预测出的结果明显优于离散时间数据，故本问主要选取附件中的“总表”内数据进行分析。首先，对总表内数据如年份、发病数、死亡数等进行相关性分析，以此选择用于预测的数据。然后，考虑到总表内数据对于时间序列预测模型并不属于高可靠性数据，故本问采用双模型并行预测——使用 BP 神经网络对每年总发病、死亡数进行预测；使用灰色马尔可夫模型对 A、B、C、D 四种方案诊断病例数进行预测。最终，分析对比两种模型的预测结果及预测误差，并选出最优值作为本问答案。

问题一的技术路线如图 2 所示：

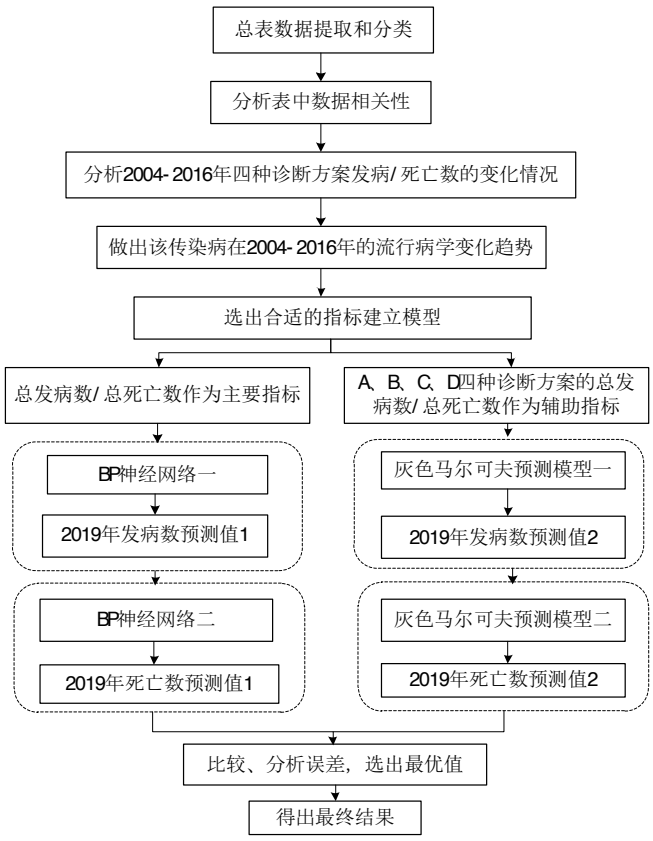

图 2 问题一技术路线图

4.2.1 年份、发病人数、死亡人数的相关性分析

相关分析是研究变量之间密切程度的一种统计方法。本问通过相关分析可以初步了解发病人数、死亡人数与时间（年份）指标间关系的密切程度进行排序，并由此可以将关系不密切的指标剔除，从而达到减少模型维数的目的。

任意两个变量间的皮尔逊相关系数 ρ 可由以下公式计算得到：

$$\rho = \frac{\sum_{k=1}^{n}(X_{ik} - \overline{X_i})(X_{jk} - \overline{X_j})}{\sqrt{\sum_{k=1}^{n}(X_{jk} - \overline{X_j})^2 \sum_{k=1}^{n}(X_{jk} - \overline{X_j})^2}}$$

其中，$\overline{X}$代表单个事件的均值；X_{ik}代表第 k 个变量第 i 个数据的数值；ρ为相关系数。X_{jk}代表第 k 个变量第 j 个数据的数值。

选用“附件 1 总表”中的“年份、总发病数、总死亡数”三类变量在 SPSS 软件中进行两两之间的相关分析，可得到各变量之间的相关系数值，表 2 为年份与总发病数之间的相关性分析结果。

表 2 “年份、总发病数、总死亡数”之间的相关性分析

		年份	发病总数	死亡总数
皮尔逊相关性	年份	1.000	-.810**	-0.275
	发病总数	-.810**	1.000	.667*
	死亡总数	-0.275	.667*	1.000
显著性 （双尾）	年份		0.001	0.363
	发病总数	0.001		0.013
	死亡总数	0.363	0.013	

**. 在 0.01 级别（双尾），相关性显著。

c. 除非另行说明，否则自助抽样结果基于 1000 个 自助抽样样本

从表 2 可以看出，显著性（双尾）为 P 值<0.05，Pearson 相关系数=-0.81 表示相关性极强，即年份与总发病数之间具有极强相关性。年份、总发病数、总死亡数三者的 Pearson 相关系数的大小可知三个因素两两之间的相关性程度由强到弱排序为：年份与总发病数，总发病数与总死亡数，年份与总死亡数。因此，本问将主要依据年份预测总发病数，再通过总发病数预测总死亡数。

4.2.2 2004-2016 年该传染病流行病学变化趋势分析

传染病发病率的趋势性检验采用 CATT(Cochran-Armitage Trend Test，CATT)方法，分析该传染病发病率随时间变化呈现出的趋势性，发病率的年递降率计算公式如下：

$$\text{年递降率} = \left(1 - \sqrt[13]{\frac{\text{2016 年该传染病发病率}}{\text{2004 年该传染病发病率}}}\right) \times 100\%$$

由上式计算可得，我国该传染病发病率的年递降率为 1.605%，有较好的控制势头。同时，根据题供数据中总表内记录的 2004-2016 年我国该传染病统计情况，可绘制出该传染病在期间的大致趋势图，如图 3 所示：

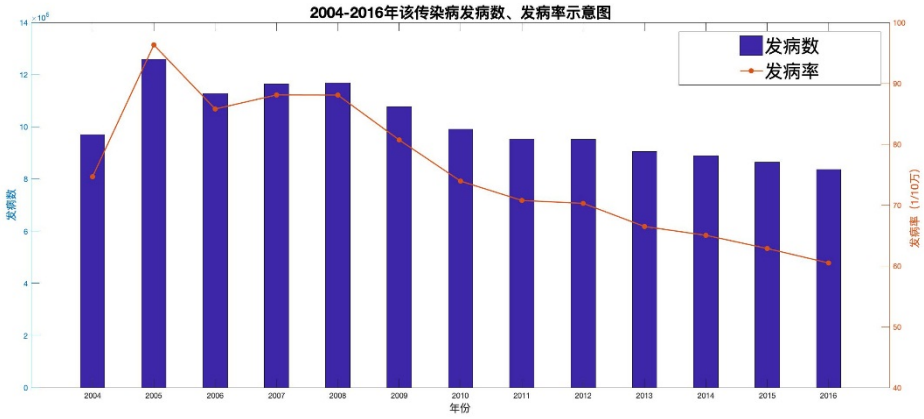

图 3 2004-2016 年全国该传染病报告病例数和发病率

图 3 较好地验证了年递降率公式，可见从 2005 年开始，全国感染该传染病的人数在稳步下降，其发病率的变化尤为明显。

同时，针对所采用的 A、B、C、D 四种诊断方案，做出了 2004-2016 年 13 年来的诊断病例数占比，以及 B、C、D 三种诊断方案诊断病例的变化趋势，如表 3、图 3 所示：

表 3　2004-2016 年四种方案诊断病例类型分布

诊断方案	发病数(例)	占比(%)
A	26356	0.20
B	2167053	16.47
C	4946028	37.59
D	6018357	45.74

由表 3 可见，A 类诊断方案诊断出的病例数量较少，故下图只使用 B、C、D 三类诊断方案进行分析比较。

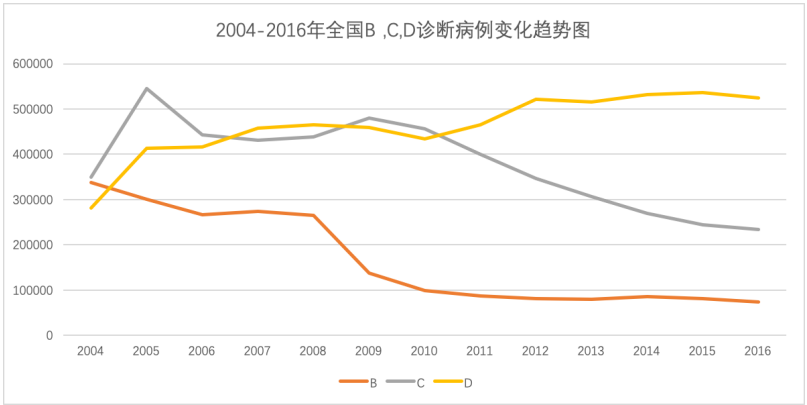

图 4　2004-2006 年三种诊断方案诊断病例变化趋势

从图 4 可看出，仅 D 类诊断方案诊断出的病例呈上升趋势，而 B、C 两类皆为下降趋势。

4.2.3 针对总发病数和总死亡数建立 BP 神经网络预测模型

4.2.3.1 BP 神经网络模型简介

BP (Back Propagation)神经网络是一种多层前馈神经网络,一般由输入层、输出层、隐藏层组成。该网络的主要特点是信号前向传递,误差反向传播。在前向传递中,输入信号从输入层经隐藏层逐层处理,到达输出层。每一层的神经元状态只影响下一层神经元的初始状态。当数据到达输出层后，将其与期望结果相比较，计算误差。并使用误差进行反向传播，从而调整神经网络中的权值和偏值，进而让使得神经网络的预测结果逼近期望结果。

在神经网络中，对于给定的输入向量$X_{k=}[x_1, x_2, \dots, x_n]$，隐藏层各神经元的激活值为：

$$S_j = \sum_{k=1}^{n}(\omega_{ij}\, x_i) + \theta_j \quad (1)$$

其中n是输入层单元数，ω_{ij}是输入层至隐藏层连接的权值，θ_j是隐藏层单元的阈值，j=1,2……p，p 是隐藏层单元数。

激活函数一般采用 sigmoid 激活函数，其表达式为：

$$\mathrm{f(x)} = \frac{1}{1 + e^{-x}}$$

将上面的激活值即下式的结果代入激活函数即可得隐含层 j 单元的输出值：

$$\mathrm{b}_j = \mathrm{f}(s_j) = \frac{1}{1 + \exp\left(-\sum_{k=1}^{n}(\omega_{ij}\, x_i) - \theta_j\right)}$$

使用上式计算每个神经元的输出值，即可得到输出层神经元的结果。接着，通过反向传播算法反复修正权值ω_{ij}和阈值θ_j，从而使得设定的损失函数最小化，即神经网络的输出结果逼近期望结果。

4.2.3.2 基于 BP 神经网络预测 2019 年全国感染该疾病的发病人数

Step1. 建立三层 BP 神经网络模型结构

BP 神经网络通常有一个或多个 sigmoid 隐藏层和线性输出层，能够对具有有限个不连续点的函数进行逼近。其学习过程由两部分组成：正向传播与反向传播。正向传播让输入信息在相应权值、阈值和激活函数的作用下传递到输出层，当输出的结果和期望值的误差大于给定精度时，则将误差反向传播。在误差返回过程中，网络修正各层的权值和阈值。如此反复迭代，最后使传递信号的误差达到允许精度。BP 神经网络模型的设计包括：网络类型及层数的确定；输入及输出变量的选择；隐藏层神经元数目的确定；激活函数的选择。本问建立的三层 BP

神经网络模型结构见下图 5。

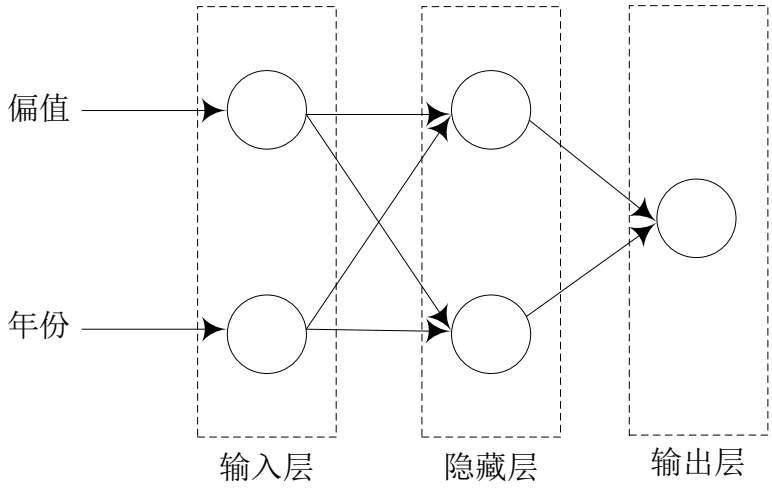

图 5　用于预测发病数的 BP 神经网络模型

Step2. 输入/输出变量的确定及其数据的预处理

由于 BP 神经网络的隐藏层一般采用 Sigmoid 转换函数，为了提高训练速度和灵敏性以及有效避开 Sigmoid 函数的饱和区，通常要求输入值在 0~1 之间。因此，首先要对输入数据进行预处理。本问使用的标准化方法如下：$P_{max} = max\{P\}$，$T_{max} = max\{T\}$，P‘ = P/P_{max}，T‘ = T/T_{max}。式中，P 为输入数据，T 为输出数据，P‘和T‘为经过归一化处理后的数据。

Step3. 神经网络拓扑结构的确定

输入层有 2 个结点，隐藏层结点数为 2，隐藏层的激活函数为 tansig；输出层结点数为 1 个，输出层的激活函数为 logsig，设置学习速率为 0.05，训练函数为 traingd，收敛误差界值为 0.005。

Step4. 基于 BP 神经网络的 2019 年发病人数预测

训练样本为全国 2004-2016 年年发病总人数。经过 125 次训练后，达到训练目标，即均方误差达到最小，得到的 2019 年该传染病发病人数预测结果如下表 4 所示，将发病人数预测结果与总发病人数真实值进行差值分析，得到总发病人数误差率。

表 4 基于 BP 神经网络的 2019 年的发病人数预测结果

年份	总发病数真实值	总发病数预测	总发病误差值	总发病误差率
2004	970279	970486	207	0.02%
2005	1259308	1259071	237	0.02%
2006	1127571	1176513	48942	4.34%
2007	1163959	1100831	63128	5.42%
2008	1169540	1171702	2162	0.18%
2009	1076938	1071766	5172	0.48%

2010	991350	1008348	16998	1.71%
2011	953275	966111	12836	1.35%
2012	951508	935153	16355	1.72%
2013	904434	911320	6886	0.76%
2014	889381	892763	3382	0.38%
2015	864015	878313	14298	1.65%
2016	836236	867028	30792	3.68%
2017		858125		
2018		850982		
2019		845134		

从上表可以看出，BP 神经网络所作出的预测值与实际值非常接近，可见模型建立较为成功。

4.2.3.3 基于 BP 神经网络预测 2019 年全国感染该疾病的死亡人数

基于 BP 神经网络预测 2019 年全国感染该疾病的死亡人数，首先建立的三层 BP 神经网络模型结构如下图 6 所示，其余分析步骤和上述一致，故不再赘述。训练样本为全国 2004-2016 年年死亡总人数。经过 125 次训练后，达到训练目标，即均方误差达到最小，得到的 2019 年该传染病死亡人数预测结果如下表 6 所示，将死亡人数预测结果与总死亡人数真实值进行差值分析，得到总死亡人数误差率。

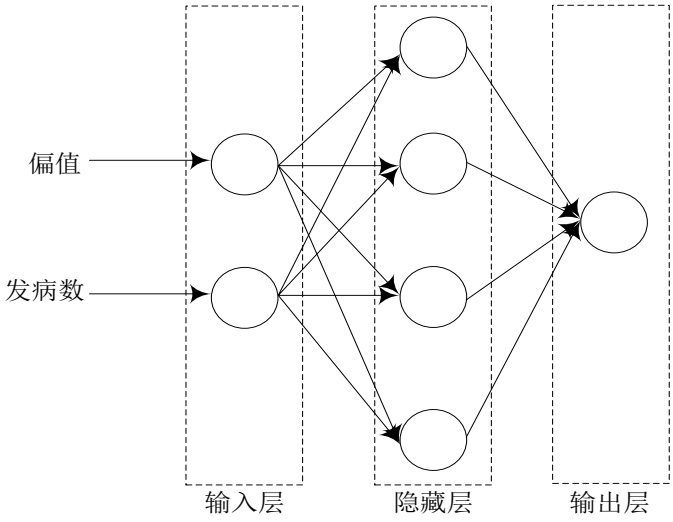

图 6　用于预测死亡数的 BP 神经网络模型

使用图 6 所示的 BP 神经网络模型根据已预测出的发病数进行训练，最终得出该传染病在 2019 年的死亡数预测值。

表 5 基于 BP 神经网络的 2019 年的死亡人数预测结果

年份	总死亡数	总死亡数预测	误差值	误差率
2004	1435	1453	18	1.27%
2005	3402	3401	1	0.04%
2006	3339	3431	92	2.76%
2007	3669	3359	310	8.44%
2008	2802	2838	36	1.30%
2009	3783	3753	30	0.79%
2010	3000	3200	200	6.66%
2011	2840	3059	219	7.73%
2012	2662	2694	32	1.20%
2013	2576	2582	6	0.22%
2014	2240	2445	205	9.14%
2015	2280	2428	148	6.49%
2016	2465	2520	55	2.24%
2017		2517		
2018		2515		
2019		2514		

从上表可以看出，BP 神经网络所作出的预测值与实际值非常接近，可见模型建立较为成功。

4.2.4 针对四种不同诊断方案下建立灰色马尔可夫预测模型

4.2.4.1 灰色马尔可夫模型

灰色马尔可夫模型预测的基本方法：首先运用灰色 GM（1,1）模型对预测的时间序列的发展趋势进行大致判断，然后用马尔可夫理论对预测结果进行精确的调整，这样就可以使预测精读大幅提高。即通过 GM（1,1）模型预测时间序列的宏观走势，根据预测值与实际值之间的相对误差，依照灰色马尔可夫理论，对相对误差进行状态划分，通过状态转移概率矩阵对灰色预测值进行微观修正。马尔可夫模型步骤如下：

Step1. 计算相对残差

相对残差为预测误差与实际值的比，即

$$\varepsilon(t) = \frac{\Delta^0(t)}{X^0(t)}$$

式中：$X^0(t)$为实际值，$\Delta^0(t)$为绝对预测误差。

Step2. 进行状态划分

相对残差$\Delta^0(t)$为一非平稳随机序列，将其划分为n个状态，任一状态区间可

表示为

$$E_i = [e_{1i}, e_{2i}]，（\mathrm{i}=1,2,\cdots,\mathrm{n}）$$

式中e_{1i}、e_{2i}为状态E_i的上下限，则相对残差的总状态集合为:

$$E = (E_1, E_{2,}, \ldots E_n)$$

状态划分的数目十分重要。状态划分应以样本数和预测的误差范围为基础，划分的数目应恰当。对于样本数较多的时间序列，数目宜多，否则状态差别不明显，失去了对波动调整的意义。如果状态太多则显得杂乱无章，难以理清头绪，故状态划分数目一般以 3~5 个为宜。

Step3. 计算状态的转移概率

状态转移概率是指客观事物由一种状态转移到另一种状态的概率。状态M_i经过m步转移到状态M_j的转移概率为

$$P_{ij}(m) = \frac{M_{ij}(m)}{M_i}$$

式中：$M_{ij}(m)$表示样本中状态E_i经过m步转移到状态E_j的个数；M_i表示状态E_i在样本中出现的次数，当M_i处于样本序列末尾时，不计入算式之中。

Step4. 计算状态转移概率矩阵

状态转移概率矩阵指系统在时刻 t 所处状态转变为时刻$(\mathrm{t}+1)$状态时条件概率矩阵，它由状态转移概率构成。随机事件的m步状态转移概率矩阵可表示为

$$\mathrm{P(m)} = \begin{bmatrix} P_{11}(\mathrm{m}) & P_{12}(\mathrm{m}) & \cdots & P_{1n}(\mathrm{m}) \\ P_{21}(\mathrm{m}) & P_{22}(\mathrm{m}) & \cdots & P_{2n}(\mathrm{m}) \\ \vdots & \vdots & \vdots & \vdots \\ P_{n1}(\mathrm{m}) & P_{n1}(\mathrm{m}) & \cdots & P_{nn}(\mathrm{m}) \end{bmatrix}$$

状态转移概率矩阵描述了系统各状态转移的全部统计规律，其各行的元素之和为 1。对象下一步状态转移方向一般通过 1 步状态转移概率矩阵即可以判断。

Step5. 确定对象转移状态

马尔可夫链的下一步与过去的状态无关，只与当前状态有关。假设预测对象处于E_k状态$(\mathrm{k}=1,2,\ldots,\mathrm{n})$，则仅考察状态转移概率矩阵第$k$行转移概率，若第j列是第$k$行中概率值最大者，则预测对象下一时刻最有可能由$E_k$状态转向$E_j$状态。若第$k$行最大元素不止 1 个，则需要计算第 2 步、3 步甚至 n 步概率矩阵，直到该元素最大值只有 1 个为止，其所处在的列即为预测对象下一步所转移的状态。

当系统满足稳定性假设时，k步状态转移概率计算公式为

$$\mathrm{P(k)} = \mathrm{P(1)}^k = \mathrm{P(1)} \cdot \mathrm{P(1)} \ldots \mathrm{P(1)}$$

式中，$\mathrm{P(1)}$为 1 步状态转移矩阵。

Step6. 确定修正后的预测值

预测值的修正与其下一步的转移状态有关。设预测对象下一步转移到E_i状态，

则灰色预测值的修正公式为

$$\hat{Y}(k)=\frac{\widehat{X^0}(k)}{1\pm 0.5\left|e_{1j}+e_{2j}\right|}$$

式中：e_{1j}、e_{2j}为E_j的上下限，预测值比实际值大者取“+”号，预测值比实际值小者取“-”号。

4.2.4.2 基于灰色马尔可夫模型的2019年发病人数与死亡人数预测

采用2004-2016年全国该传染病在A、B、C、D四种不同诊断方法下的发病人数和死亡人数数据，建立GM（1,1）模型，并对未来3年（2017年，2018年，2019年）进行预测，结果见下表6所示，平均相对误差为2.15%，模型的精确度较高。

表6 基于灰色马尔可夫模型的2019年发病人数预测结果

年份	总发病数真实值	总发病数预测	总发病数误差	总发病数误差率
2004	970279	970279	0	0.00%
2005	1259308	1275705	16397	1.30%
2006	1127571	1203043	75472	6.69%
2007	1163959	1160474	3485	0.30%
2008	1169540	1130515	39025	3.34%
2009	1076938	1039215	37723	3.50%
2010	991350	990670	680	0.07%
2011	953275	940608	12667	1.33%
2012	951508	924820	26688	2.80%
2013	904434	912078	7644	0.85%
2014	889381	885473	3908	0.44%
2015	864015	836553	27462	3.18%
2016	836236	870883	34647	4.14%
2017		849840		
2018		841504		
2019		835515		

同理，基于灰色马尔可夫模型对2019年死亡人数进行预测，得到的结果如下表7所示。

表 7 基于灰色马尔可夫模型的2019年死亡人数预测结果

年份	总死亡数	总死亡数预测	总死亡数误差	总死亡数误差率
2004	1435	1435	0	0.00%
2005	3402	3471	69	2.02%

2006	3339	3114	225	6.75%
2007	3669	3155	514	14.00%
2008	2802	2594	208	7.42%
2009	3783	3352	431	11.40%
2010	3000	3166	166	5.52%
2011	2840	2690	150	5.29%
2012	2662	2568	94	3.53%
2013	2576	2352	224	8.71%
2014	2240	2239	1	0.04%
2015	2280	2313	33	1.43%
2016	2465	2429	36	1.48%
2017		2328		
2018		2329		
2019		2333		

从上表可见，在 2007 年和 2009 年时，灰色预测模型存在较大的预测误差值，但在近些年来，误差明显收敛，这是因为之前数据量过小，年份与死亡数的数据过少，灰色马尔可夫在面对长数据时有着更好的表现。

4.2.5 两个预测模型的效果评价

预测结果评价方法选用平均相对误差及标准差来评价模型的精确度。它不受时间单位的影响，可以比较不同的时间序列预测值的有效性，通常平均相对误差应控制在 5%以内，其值越接近 0 准确率越高；标准差越小，说明预测结果的误差数的离散程度越小，效果越好。因此，根据表 8 中 BP 神经网络模型与灰色马尔可夫模型预测结果误差率的均值和标准差结果可以看出，BP 神经网络的误差最小，效果最好；灰色马尔可夫模型的效果次之。

表 8 两个模型预测结果的均值和标准差

评价指标	总发病数误差率比较		总死亡数误差率比较	
	BP 神经网络模型	灰色马尔可夫模型	BP 神经网络模型	灰色马尔可夫模型
均值	1.67%	2.15%	3.71%	5.20%
标准差	0.017	0.019	0.033	0.042

4.2.6 最终结果确定

由于 BP 神经网络模型对全国感染该疾病的发病人数和死亡人数预测效果均优于灰色马尔可夫模型，故以 BP 神经网络的预测结果为最终结果，得到预测的 2019 年全国感染该疾病的发病人数和死亡人数为下表 9 所示。

表 9 基于 BP 神经网络模型预测 2019 年的发病人数和死亡人数结果

年份	总发病数真实值	总发病数预测	总死亡数	总死亡数预测
2004	970279	970486	1435	1453
2005	1259308	1259071	3402	3401
2006	1127571	1176513	3339	3431
2007	1163959	1100831	3669	3359
2008	1169540	1171702	2802	2838
2009	1076938	1071766	3783	3753
2010	991350	1008348	3000	3200
2011	953275	966111	2840	3059
2012	951508	935153	2662	2694
2013	904434	911320	2576	2582
2014	889381	892763	2240	2445
2015	864015	878313	2280	2428
2016	836236	867028	2465	2520
2017		858125		2517
2018		850982		2515
2019		845134		2514

4.3 问题二模型的建立与求解

问题二要求结合 2004 年,2007 年,2010 年,2013 年和 2016 年这五年按不同地区和职业分类统计的数据，建立该传染病模型，并且预测 2019 年传染病防控排名前 3 位的重点区域和重点人群。首先，针对这五年按不同地区和职业分类统计的该传染病的发病人数和死亡人数数据，应用地理信息软件 Arcgis10.4 进行全部和局部空间自相关分析，探测该传染病发病的时空聚集；然后依据这五年该传染病的时空分布特征，建立基于自回归求和移动平均（ARMIA）的该传染病传播模型；最后通过该模型，预测 2019 该传染病防控的区域和职业人群并进行排序，选择排名前 3 位的重点区域和重点人群。

问题二的技术路线如图 7 所示：

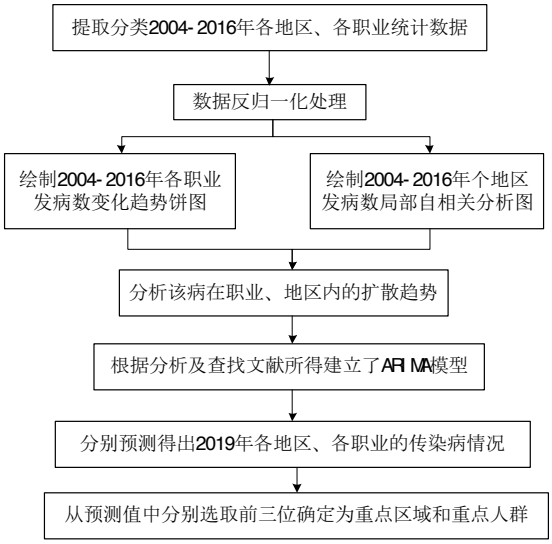

图 7 问题二技术路线图

4.3.1 年均发病率空间分布

2014-2015 年全国该传染病年均发病率采用 Jenks 分组方法实现可视化主题地图展示，ArcGIS 软件中自然断点分级方法（Jenks）是基于数据固有的自然分组，反映的是“物以类聚”的思想，是 GIS 绘制分级主题图常用方法。该方法采用尽可能的方式对空间区域某种属性值分组，依据各组 SSD（Sum of Squared Differences）最小的分割点划分等级，该方法能够识别出分类间隔，使得数据分类后各类别内部的差异最小化，类别之间的差异最大化，以便于观察空间区域某种属性值（如：该传染病的年均发病率）的分布差异。SSD 计算公式如下：

$$SSD_{i,j} = \sum_{n=i}^{j} (A[n] - \overline{X_{i,j}})^2 = \sum_{n-i}^{j} A[n]^2 - \frac{\left(\sum_{n=i}^{j} A[n]\right)}{j-i+1}$$

式中，$1 \le i < j < n$，n 为空间区域数量，A 为按照顺序排列的某一个数据组，$\overline{X_{i,J}}$为第 i 个空间区域属性值和第 j 个空间区域属性值的平均值。

4.3.2 该传染病发病率的空间自相关分析

结合 2004 年，2007 年，2010 年，2013 年和 2016 年各年全国该传染病的发病数和死亡人数，计算各省该传染病的年均发病率，并将数据导入地理信息软件 Arcgis10.4 进行全部和局部空间自相关分析和时空扫描分析，分为全局自相关分析和局部自相关分析，由于缺少南沙群岛及其阳面数据，故本问导入的地图中不包含该区域。全局自相关是研究该传染病统计发病率在全国范围内是否具有空间聚集性。以 Moran’s I 作为统计指标，其取值范围介于-1~1 之间。取值为正，表示数据呈正相关，越接近 1，数据的正自相关性越强，地域聚集性越高。反之，数据的负相关性越强，数据越分散。取值月接近于 0，数据则越可能是随机分布。局部自相关通过局部 Moran 指数来探测研究区域内研究指标的具体聚集范围。

其聚集类型可分为 4 种：高高值聚集；低低值聚集；高低值聚集；低高值聚集。

4.3.2.1 全局自相关分析

Moran's I 系数是全局自相关分析的常用指标，由 Moran 于 1950 年提出，从整体上描述研究区域的空间对象或者属性值的空间分布状态，全局 Moran's I 系数计算公式如下：

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\left[\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\right]\sum_{i=1}^{n}(x_i - \bar{x})^2}, (i \neq j)$$

式中，n 为空间单元个数，即 31 个省级行政单位，w_{ij}为区域 i 和 j 的空间权重系数，x_i和x_j分别为区域 i 和 j 的该传染病发病率，$\bar{x}$为 31 个省级行政单位该传染病年平均发病率。

通过计算 2004 年，2007 年，2010 年，2013 年和 2016 年历年的全局自相关指数 Moran's I 均大于 0，且 P 值均小于 0.05，差异具有统计学意义，具体见下表 10。

表 10 2004, 2007, 2010, 2013 和 2016 年中国大陆该传染病发病率的全局自相关分析

年份	Moran's I	V(I)值	Z 值	P 值
2004	0.34	0.006	4.807	＜0.001
2007	0.381	0.0057	5.464	＜0.001
2010	0.4	0.0057	5.718	＜0.001
2013	0.367	0.0057	5.313	＜0.001
2016	0.364	0.0056	4.832	＜0.001

从表 10 可以看出，该传染病发病率的地区分布层空间正自相关，存在空间聚集性，即存在高发病区和低发病区。

4.3.2.2 局部自相关分析

全局 Moran's I 系数只能从总体上反映空间区域某种属性值（该传染病发病率）的空间自相关关系，无法反映空间区域某种属性值聚集区域的聚集类型及位置。在分析空间区域某种属性值的空间自相关关系时，如充分考虑空间异质性的存在，会出现下面两种情形：

（1）部分局部区域空间某种属性值的自相关系数有统计学意义，才使得全局自相关分析有统计学意义；

（2）区域空间某种属性值的全局空间自相关系数无统计学意义，但某些局部区域存在统计学意义。

Anselin 在 1995 年提出了局部空间自相关性分析统计量（Local Indicators of Spatial Auto-correlation，LISA），它需要满足以下两个条件：

（1）每个局部空间区域的 LISA 值反映的是自身某种属性值与邻近空间区域某种属性值的空间聚集性指标；

（2）局部空间区域的 LISA 值相加与全局空间自相关性指标成正比。

整体上看，2004-2016 年，全国大陆该传染病发病的热点区域集中西部地区。2004 年-2007 年，西南部和南部的贵州、广西和海南为发病的高高值聚集区。2010 年以后，新疆和西藏的该传染病疫情迅速上升，称为新的高发病区，并且始终保持较高的水平。海南的该传染病疫情则逐渐减轻。东部地区是该传染病发病的冷点区域。其中北京、天津、山东、上海是稳定的低低值聚集区，该传染病发病率始终处于较低水平。结果见下图。

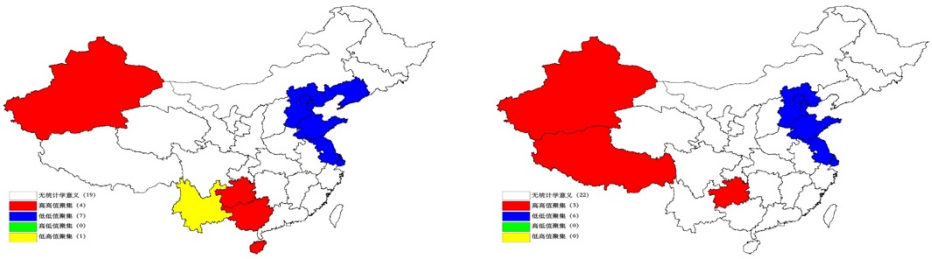

图 8 2004-2007（左）2007-2010（右）年中国大陆该传染病发病率的局部自相关分析

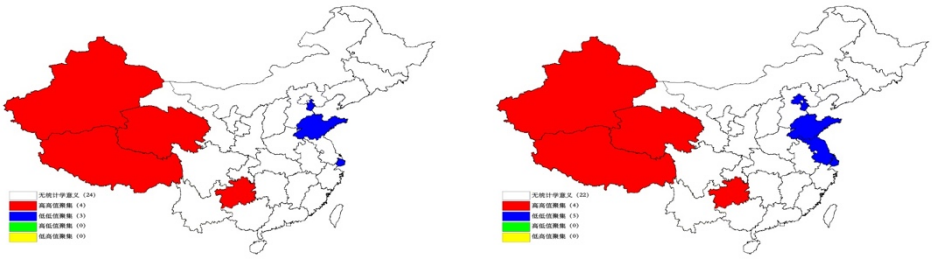

图 9 2010-2013（左）2013-2016（右）年中国大陆该传染病发病率的局部自相关分析

4.3.3 2004-2016 年全国该传染病的职业分布

根据 2004 年，2007 年，2010 年，2013 年和 2016 年所给数据中该传染病患者职业分布显示，农民职业患者人数最多，每年占比均在 60%以上，其次是家务及待业人数，保育员及保姆患病人数最少，见图 10、表 11。可以看到，农民和家务及待业人数占数据总发病人数的比例逐年增加，学生、工人和民工占比也逐年上升。

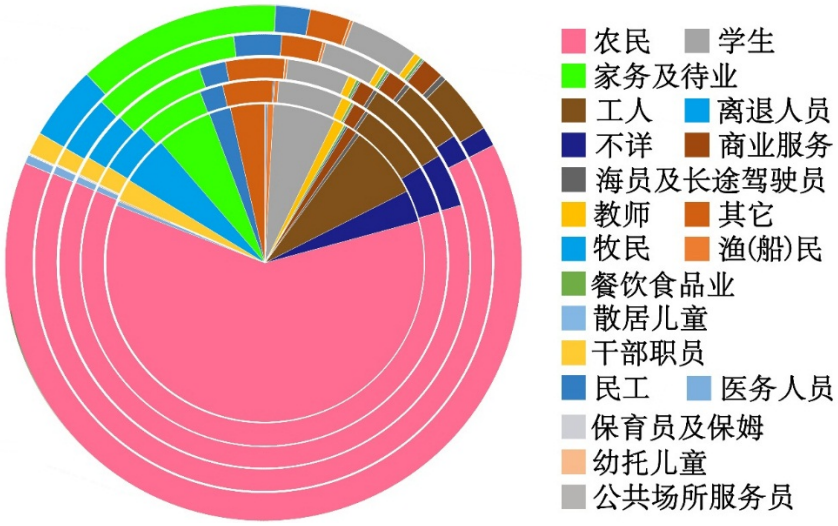

图 10 2004, 2007, 2010, 2013, 2016 年全国该传染病患者职业占比情况

表 11 不同职业在各年的占比情况

职 业	2004	2007	2010	2013	2016
幼托儿童	0.23%	0.13%	0.04%	0.02%	0.02%
散居儿童	0.64%	0.35%	0.18%	0.19%	0.18%
学生	6.35%	5.77%	4.94%	4.10%	4.32%
教师	1.07%	0.85%	0.64%	0.48%	0.43%
保育员及保姆	0.02%	0.01%	0.01%	0.01%	0.01%
餐饮食品业	0.22%	0.22%	0.23%	0.23%	0.19%
公共场所服务员	0.19%	0.08%	0.11%	0.11%	0.09%
商业服务	1.01%	1.16%	1.22%	1.39%	1.39%
医务人员	0.39%	0.30%	0.31%	0.33%	0.38%
工人	7.44%	6.58%	6.93%	4.61%	3.80%
民工	3.24%	4.25%	4.21%	1.82%	1.21%
农民	60.11%	60.94%	60.96%	63.90%	63.81%
牧民	0.43%	0.48%	0.53%	0.58%	0.55%
渔(船)民	0.11%	0.08%	0.07%	0.05%	0.05%
海员及长途驾驶员	0.09%	0.09%	0.11%	0.08%	0.07%
干部职员	2.17%	1.71%	1.48%	1.26%	1.38%
离退人员	4.95%	4.45%	3.98%	4.11%	4.66%
家务及待业	5.66%	6.04%	7.29%	10.45%	12.75%
不详	2.18%	2.04%	2.10%	3.38%	2.18%
其它	3.51%	4.45%	4.66%	2.91%	2.54%

4.3.4 建立 ARMIA 传染病传播模型

ARIMA 模型即自回归求和移动平均模型,记为 ARIMA(p,d,q),其中 p 和 q 为自动回归和移动平均阶数,d 为差分阶数。基本思想是利用时间序列所具有的自相关性表征预测对象发展的延续性,这种自相关性一旦被相应的数学模型描述出来,就可以从时间序列的过去值及现在值预测其未来的值[5]。ARIMA 模型建模步骤:①进行模型识别;②估计模型参数;③模型检验;④预测[6]。

由于按职业分类的传染病统计数据总，农民这一职业所占比例最大，故以职业为农民的数据为例说明使用 ARMIA 模型预测 2019 传染病发病数的建模流程。具体建模步骤如下：

Step1. 模型识别

ARIMA(p,d,q)模型中的 d 即平稳化过程的差分阶数,需分析时间序列的平稳性来得到。可通过单位根(ADF)检验来验证时间序列平稳性,使用 SPSS Statistics 24 对序列经进行检验，结果为非平稳序列(t=-3.32,P>0.05),进行一阶差分后序列平稳化(t=-5.65,P<0.05),无须二阶差分,所以 d 取值为 1,模型初步确定为 ARIMA(p,1,q)。

Step2. 估计模型参数

为确定 p、q 取值,需计算 p、q 不同取值下该模型的 AIC(Akaike 信息准则)值和正态化 BIC(Bayesian 信息准则)值,这两个数值越小,ARIMA 模型拟合优度和简约性越好。使用 SPSS Statistics 24 分别计算不同 p,q 取值下模型的 AIC 和正态化 BIC 值。结果如表 12 所示。可知 ARIMA(0,1,1)模型正态化 BIC 值最小,为 22.74,AIC 值最小,为 823.83。故选取 ARIMA(0,1,1)模型。

表 12 ARIMA 模型拟合优度检验

模型	AIC	正态化 BIC
ARIMA(0,1,1)	823.83	22.74
ARIMA(1,1,0)	841.41	23.41

注:p、q 取值≤2 。

Step3. 模型检验

对 ARIMA(0,1,1)模型进行检验，得到时间序列残差自相关及偏自相关系数图，见下图 11。

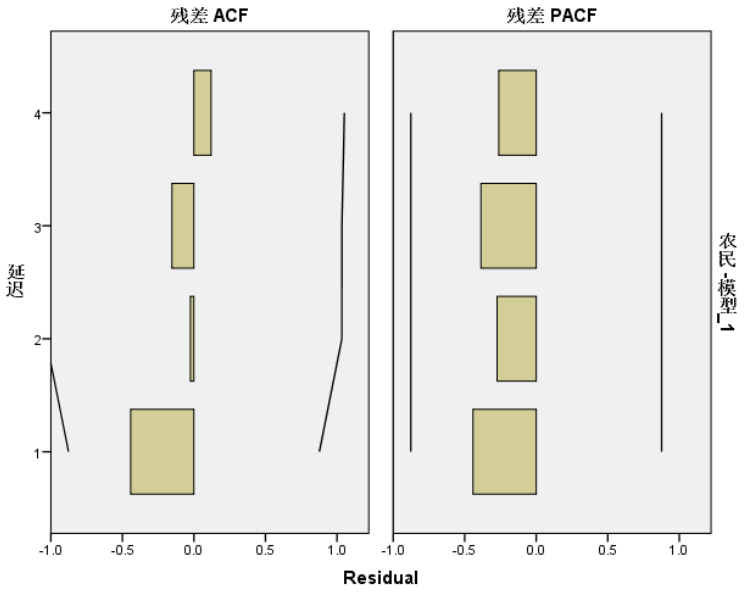

图 11 时间序列残差自相关及偏自相关系数图

从图 11 中可以看出,模型残差序列的自相关系数(ACF)和偏自相关系数(PACF)均在 95% CI 内。验证了残差已是随机分布的白噪声序列,所选 ARIMA(0,1,1)模型合适。

Step4. 发病人数预测

用 ARIMA(0,1,1)模型拟合农民 2004-2016 年因患有某传染病发病人数，即得到 2019 年传染病农民的发病人数为 524125。

4.3.5 基于 ARMIA 传染病传播模型预测 2019 年防控重点区域和人群

4.3.5.1 基于 ARMIA 传染病传播模型预测 2019 年防控重点人群

利用上述 ARIMA 方法预测 2019 年中国大陆不同职业分类的该传染病防控重点人群，所使用 ARIMA 模型的参数及预测结果，得到图 12 为 2019 年不同职业患该传染病的占比图。

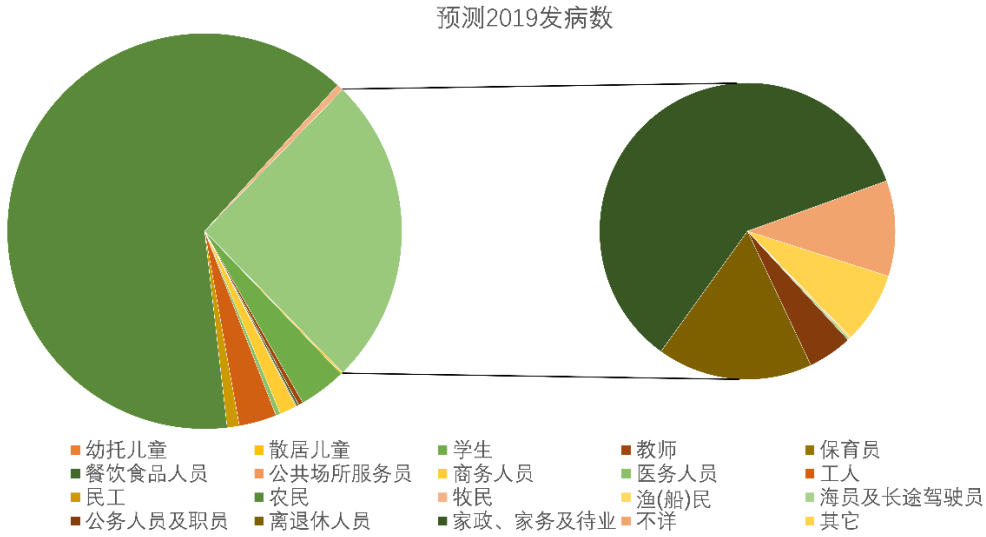

图 12 2019 年该传染病发病人数预测值分布（按不同职业划分）

对预测的 2019 年不同职业发病数进行排序后的结果如表 13 所示。对应的时间序列残差自相关及偏自相关系数图，详见附录 1。

表 13 基于 ARMIA 模型预测 2019 年不同职业该传染病的发病数

职 业	2004 发病数	2007 发病数	2010 发病数	2013 发病数	2016 发病数	ARIMA 模型参数	预测 2019 发病数
农民	583238	709324	604333	577932	533637	（1,1,1）	524125
家务及待业	54916	70294	72289	94520	106593	（0,0,2）	124401
离退人员	47991	51795	39442	37154	38959	（0,1,1）	35489
学生	61578	67157	48961	37040	36094	（0,0,2）	31885
工人	72143	76635	68722	41681	31779	（2,0,0）	25337
不详	21189	23717	20802	30549	18211	（0,0,1）	21799
其它	34028	51829	46178	26332	21239	（0,0,2）	16222
商业服务	9797	13451	12053	12536	11611	（0,0,1）	11193
干部职员	21032	19862	14628	11366	11530	（0,0,2）	10128
民工	31404	49512	41758	16440	10125	（0,0,2）	8153
牧民	4156	5644	5263	5204	4599	（0,0,2）	4432
教师	10396	9928	6341	4350	3560	（0,1,1）	3141
医务人员	3803	3510	3101	3017	3170	（1,0,1）	3140
餐饮食品业	2175	2558	2328	2049	1620	（1,0,1）	1482
散居儿童	6179	4097	1821	1757	1493	（2,0,0）	1112
公共场所服务员	1821	987	1087	996	737	（1,0,1）	772
海员及长途驾驶员	900	1079	1041	724	580	（0,0,0）	500
渔(船)民	1113	981	672	496	448	（0,0,2）	395
幼托儿童	2239	1490	437	203	189	（0,1,0）	186
保育员及保姆	181	109	93	88	62	（0,0,2）	58

从上表可以看出，2019 年该传染病防控排名前 3 位的重点人群分别为：农民，家务及待业，离退人员。

4.3.5.2 基于 ARMIA 传染病传播模型预测 2019 年防控重点区域

利用 4.3.3 中所述的 ARIMA 方法预测中国大陆各地区 2019 年因患有该传染病发病人数，所使用 ARIMA 模型的参数及预测结果如表 14 所示。对应的时间序列残差自相关及偏自相关系数图，见附录 2。

表 14 基于 ARMIA 模型预测 2019 年不同地区该传染病的发病数

地区	2004（/10 万人）发病数	2007（/10 万人）发病数	2010（/10 万人）发病数	2013（/10 万人）发病数	2016（/10 万人）发病数	ARIMA 模型	2019（/10 万人）发病数
西藏	70.33	88.03	114.40	136.35	151.09	（1,0,1）	182.91

新疆	134.60	195.69	162.48	170.35	182.72	（1,0,1）	182.73
青海省	71.86	87.77	86.47	104.76	127.71	（0,0,2）	145.67
贵州省	119.69	175.78	140.99	133.59	129.72	（0,0,2）	114.76
广西	118.81	128.51	102.19	96.22	85.52	（1,0,1）	73.72
黑龙江	101.16	106.06	95.66	88.22	80.43	（0,0,1）	73.24
海南省	122.64	142.09	109.03	94.37	83.61	（1,0,0）	70.18
湖南省	64.02	103.93	86.38	88.31	74.99	（1,0,1）	68.94
江西省	87.57	93.72	84.52	71.85	71.57	（0,0,2）	66.12
湖北省	104.85	112.31	84.64	82.85	74.28	（1,0,0）	63.11
广东省	66.15	98.72	91.86	72.82	70.84	（0,0,2）	62.66
重庆市	142.88	124.80	89.09	82.61	72.58	（0,0,2）	60.53
四川省	91.60	103.31	83.16	75.06	65.20	（0,1,1）	55.80
辽宁省	54.46	56.59	59.33	59.64	51.45	（0,0,1）	53.21
云南省	52.93	61.69	55.23	56.48	55.13	（0,0,1）	52.59
河南省	68.84	96.06	73.75	67.01	59.80	（0,1,1）	49.60
吉林省	84.23	72.38	83.56	62.02	50.13	（0,0,1）	48.90
陕西省	90.90	84.71	68.58	60.76	56.01	（0,0,2）	47.50
安徽省	55.26	88.52	66.37	62.46	56.29	（0,0,0）	45.81
甘肃省	78.45	121.89	93.31	69.14	57.89	（1,0,1）	44.71
河北省	55.70	63.41	57.77	48.98	45.06	（0,0,0）	39.57
浙江省	87.16	83.34	60.75	52.80	48.33	（0,0,2）	38.70
内蒙古	77.40	89.38	72.58	57.68	48.13	（1,0,1）	38.46
山西省	54.79	70.56	66.73	52.34	38.46	（0,0,2）	35.05
宁夏	65.63	68.97	56.68	44.13	39.61	（1,0,1）	32.19
福建省	82.42	82.49	57.88	46.89	42.36	（0,1,1）	31.98
江苏省	64.12	61.56	52.95	44.02	35.82	（0,0,2）	30.61
山东省	33.38	43.46	41.96	36.96	30.47	（0,0,1）	28.66
北京市	47.98	49.02	40.88	35.12	30.98	（1,1,0）	25.96
上海市	36.97	33.92	28.64	28.84	27.23	（2,0,0）	24.90
天津市	28.38	34.55	24.15	21.39	20.94	（0,0,2）	16.47

从上表可以看出，2019 年该传染病防控排名前 3 位的重点地区分别为：西藏，新疆和青海省。

4.4 问题三模型的建立与求解

问题三要求结合地区经济发展的相关公开数据，选择一个角度尝试建立该传染病与经济发展的数学模型，并分析得到结果。本问在问题一和问题二的基础上，

通过查阅大量文献和资料，以广东省为例，查找了该省 2004、2007、2010、2013、2016 年共 5 年的人口结构、地区生产总值、城镇居民可支配收入、地方政府文教卫生支出、每 10 万人拥有卫生从业人员数等数据，分析数据相关性。然后，根据相关性结果选取出“流动人口”、“地方政府医疗支出”、“城镇化水平”、“居民人均可支配收入”四项指标，并提出 “可获得医疗帮助的能力”这一新的角度，以此建立多元线性回归模型。最后，使用该模型在广东省这 5 年内进行纵向对比，并与 2016 年上海传染病情况进行横向对比，以此来验证模型的可行性和可靠性。

问题三的技术路线图如图 13 所示：

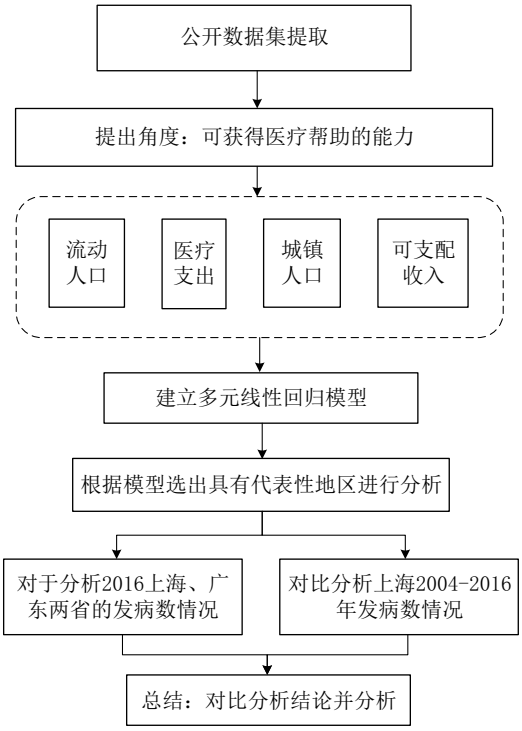

图 13 问题三的技术路线图

4.4.1 传染病与地区经济发展相关指标数据确定

本问以上海市为例，选取地区 2004 年，2007 年，2010 年，2013 年和 2016 年“常住人口、城镇人口、地区生产总值、人均可支配收入、医疗支出、每 10 万人拥有卫生从业人员、流动人口等 7 个传染病与地区经济发展相关指标数据，具体数据见表 15。其中“常住人口、人均可支配收入（以下简称为“可支配”）、医疗支出、卫生员数、城镇人口、GDP”这 6 个指标数据来源于《国家统计年鉴》，发病数为原题所给数据。

表 15 上海市 2004-2007 年地区经济发展相关指标数据

年份	2004	2007	2010	2013	2016
常住人口（/10 万人）	183.5	206.4	230.3	241.5	242.0
可支配（万/10 万人）	148710.73	220997.04	299277.73	397852.09	508223.27
卫生员数/10 万人	130.80	155.80	171.90	192.30	217.10

医疗支出(万/10 万人）	2453.13	4303.84	6950.53	8899.40	15830.58
城镇人口（/10 万人）	81160	88700	89304	89600	87900
GDP（万元）	8072.83	12494.01	17165.98	21818.15	28178.65
流动人口(/10 万人)	26299.47	33181.17	38665.72	40693.54	67581.55

其中流动人口是指离开户籍所在地的县、市或者市辖区，以工作、生活为目的异地居住的成年育龄人员，并且人口流动主要是由农村流向城市，由经济欠发达地区流向经济发达地区，由中西部地区流向东部沿海地区。由于流入人口是该地区非户籍人口，而户籍与教育、医疗、社会保障等挂钩，离开了户籍所在地流动人口就权益得不到保障，故流动人口在传染病的防治方面获得的保障较低。依据《上海统计年鉴 2018》得到上海省各年末户籍总人口数；依据《国家统计年鉴》得到上海省常住人口数，并通过流动人口的计算公式，即：流动人口数（万人）=常住人口（万人）-年末户籍总人口（万人），计算得到广东省各年流动人口数（万人），具体如表 16 所示。

表 16 上海流动人口计算结果

年份	年末户籍总人口/万人	常住人口/万人	流动人口/万人
2004	1352.39	1834.98	482.59
2007	1378.86	2063.58	684.72
2010	1412.32	2302.66	890.34
2013	1432.34	2415.15	982.81
2016	1450.00	2419.70	969.70

由于本问后续分析采用反归一化数据，即各数据的量纲统一为“/10 万人”，其中流动人口（/10 万）=流动人口数/常住人口*100000，结合表 10 流动人口（/万人）计算得到表 16 中流动人口（/10 万人）数据。

通过 person 相关性分析可知，常住人口指标与可支配指标相关（person=0.917,P<0.05），卫生员数指标与医疗支出指标相关（person=0.965,P<0.05），GDP 指标与可支配指标相关（person=0.999,P<0.05），故本文剔除常住人口、卫生员数、GDP 指标，仅选取可支配、医疗支出、城镇人口作为上海市经济发展指标，具体数据如下表 17 所示。

表 17 上海经济发展指标数据选取结果

年份	2004	2007	2010	2013	2016
可支配（万/10 万人）	148710.7	220997	299277.7	397852.1	508223.3
医疗支出(万/10 万人）	2453.134	4303.842	6950.53	8899.395	15830.58
城镇人口（/10 万人）	81160	88700	89304	89600	87900

流动人口(/10 万人)	26299.47	33181.17	38665.72	40693.54	67581.55
发病数（/10 万人）	36.97003	33.92442	28.63656	28.84058	27.22727

4.4.2 建立多元线性回归模型分析经济发展指标与传染病的关系

4.4.2.1 多元线性回归数学模型介绍

设随机变量 y 随着 m 个自变量x_1，x_2，…，x_m变化，并有如下线性关系式：

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \varepsilon \quad (1)$$

此式称为回归方程，其中β_0，β_1，…，β_m称为回归系数，是$m+1$个待估计参数，ε是随机变量（剩余参数）。

回归分析的主要问题是根据x_1，x_2，…，x_m，y的N组观测数据$(x_{k1}$，…，x_{km}，$y_k)k=1$，2，…，N。给出各回归分析系数β_i的估计值$\widehat{\beta_i}$，同时对$\widehat{\beta_i}(i=0$，1，2，…，m)各作统计检验，以便说明估计值的可靠性。

将观测数据带入回归方程（1）得到如下结构式：

$$\begin{cases} y_0 = \beta_0 + \beta_1 x_{11} + \cdots + \beta_m x_{1m} + \varepsilon_1 \\ y_N = \beta_0 + \beta_1 x_{N1} + \cdots + \beta_m x_{Nm} + \varepsilon_N \end{cases} \quad (2)$$

其中ε_1，…，ε_N是 N 个独立且服从同一正太分布N(0，σ)的随机变量

假设$Y=\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$，$X=\begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nm} \end{bmatrix}$，$\beta=\begin{bmatrix} \beta_0 \\ \vdots \\ \beta_m \end{bmatrix}$，$\varepsilon=\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix}$，则（2）式对应的矩阵方程：

$$Y = X\beta + \varepsilon \quad (3)$$

根据公式（3），结合回归系数的最小二乘估计求的β的估计值$\hat{\beta}$，即可得到多元线性回归方程。

4.4.2.2 基于多元线性回归模型的经济发展指标与传染病关系分析

Step1. 回归拟合中的输入指标变量和去除变量

运用 SPSS 软件将“可支配、医疗支出、城镇人口、流动人口”4 个指标作为输入变量进行多元线性回归分析，得到结果如下表 18 所示。

表 18 多元线性回归输入和除去变量结果

模型	输入的变量	除去的变量	方法
1	流动人口(/10 万人), 城镇人口（/10 万人）, 可支配（万/10 万人）, 医疗支出(万/10 万人）[b]		输入

a. 因变量：发病数（/10 万）

b. 已达到“容差 = .000”限制。

从表 18 中可知，所建立的模型中包含的流动人口、城镇人口、可支配、医疗支出变量参与了回归拟合，未剔除变量。

Step2. 方差分析

回归拟合中的方差分析结果如表 19 所示，从表中可以看出回归平方和为 68.60，残差平方和趋近于 0,。回归平方和占总计平方和的绝大部分，说明线性模型解释了总平方和的绝大部分，模型拟合效果较好。

表 19 回归拟合过程中的方差分析结果

	平方和	自由度	均方
回归	68.604	4	17.151
残差	0.000	0	
总计	68.604	4	

a. 因变量：发病数（/10 万）

b. 预测变量：(常量), 流动人口（/10 万）, 城镇人口（/10 万人）, 可支配（万/10 万人）, 医疗支出(万/10 万人）

Step3. 回归方程系数估计

回归方程系数估计值计算结果如表 20 所示，其中 Beta 是标准化回归系数，是所有的变量按统一方法标准化后拟合的回归方程中各标准化变量的系数，具有可比性。由表 14 可知，居民人均可支配收入的标准化回归系数是 4 个变量中最大的，即居民人均可支配收入对发病人数的影响最大。

表 20 模型的回归系数的估计值

	未标准化系数		标准化系数	相关性			共线性统计
	B	标准误差	Beta	零阶	偏	部分	容差
(常量)	87.619	0.000					
可支配（万/10 万人）	2.008E-04	0.000	6.896	-0.907	1.000	0.268	0.002
医疗支出(万/10 万人）	-1.208E-02	0.000	-15.103	-0.848	-1.000	-0.301	0.000
城镇人口（/10 万人）	-1.323E-03	0.000	-1.121	-0.766	-1.000	-0.382	0.116
流动人口(/10 万人)	2.147E-03	0.000	8.153	-0.787	1.000	0.302	0.001

a 因变量：发病数（/10 万）

根据表 20，得到广东省发病人数的多元线性回归方程如下：
当年发病数=a+b×可支配+c×医疗支出+d×城镇人口+e×流动人口，其中 a、b、c、d、e 为改模型的系数，如表 21 所示：

表 21 多元线性回归模型系数值

系数	系数值
a	2.008×10-4
b	-1.2.8×10-2
c	-1.323×10-3
d	2.147×10-3

4.4.2.2 多元线性回归模型的纵向对比

使用上节所得的多元线性回归方程拟合 2004-2016 年上海地区传染病数，结果见表 21，并由此绘制折线图，如图 14 所示。

表 21 2004-2016 年上海地区传染病回归模型数据

年份	真实值	回归值
2004	36.97	37.54
2007	33.92	34.38
2010	28.64	28.76
2013	28.84	28.44
2016	27.23	27.50

图 9 2004-2016 年上海地区传染病发病数回归模型示意图

由上述可知，使用多元线性回归方程拟合的回归值与题中所给的真实值差别不大，这说明多元线性回归方程很好的拟合了表 21 的数据。

4.4.2.2 多元线性回归模型的横向对比

为了更好地验证多元线性回归模型的可靠性，本文选用 2004-2016 年广东省经济发展指标数据选取，如表 22 所示。其中流动人口数据由《广东统计年鉴 2017》中的数据计算得出，“人均可支配收入（以下简称为“可支配”）、医疗支出、城镇人口”这 3 个指标数据来源于《国家统计年鉴》，发病数为原题所给数据。

表 22 广东经济发展指标数据选取结果

年份	2004	2007	2010	2013	2016
可支配（万/10 万人）	97710.45	132483.59	184840.47	194381.25	254344.51
医疗支出(万/10 万人）	800.24	1457.24	2912.00	5348.78	10199.38
城镇人口（/10 万人）	58360.00	63140.00	66180.00	67760.00	69200.00
流动人口(/10 万人)	16732.25	18433.80	22523.95	21514.34	20012.22
发病数（/10 万人）	66.15	98.72	91.86	72.82	70.84

使用上节中求得的多元线性方程拟合 2004-2016 年广东地区传染病数，结果见表 23，并由此绘制折线图，如图 15 所示。

表 23 2004-2016 年广东省传染病回归模型

年份	真实值	回归值
2004	66.15	66.21
2007	98.72	98.78
2010	91.86	91.92
2013	72.82	72.83
2016	70.84	70.77

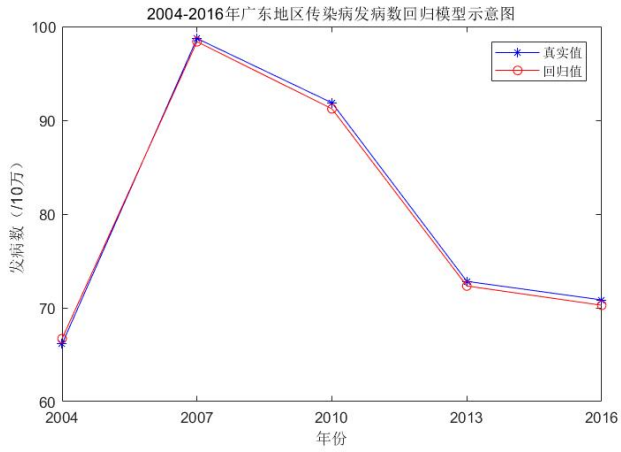

图 15 2004-2016 年广东省回归模型方程拟合

由上述可知，使用多元线性回归方程拟合的回归值与题中所给的真实值差别不大，这说明多元线性回归方程很好的拟合了表 23 的数据。

4.5 问题四的求解

据题意可知，问题四需要结合前述问题分析结果，给卫生健康委员会相关部

门写一封公开信，并给出对该传染病疫情防治的看法和建议。首先，对已完成问题和已有数据进行分析研判，由于农民、学生、离退人员、家政及待业等人群中发病数较多，且新疆、西藏、青海等地区发病率较高，故大致判断出该传染病应出现在抵抗力较弱、不注重平时保养、身体出现小问题不在意等群体身上，且较大可能出现在经济欠发达、地广人稀、医务人员稀少或流动人口过多的地区。其次，分析其他经济指标，发现有些政策以及政府地方财政等效果不明显，本问将主要针对这一情况作出相应分析和判断。最后，结合本文预测结果和分布趋势，给出对该传染病疫情防治的看法和建议。

根据表 24 所示内容可看出，该传染病主要发病人群集中在学生、农民、离退人员、家务及待业等群体中，上述人群中学生和离退人员由于年纪较小或较大，身体素质不强、缺乏体育锻炼、抵抗力欠缺等因素导致其成为该传染病的高发人群；而农民和家务及待业人群，大多可能没有非常在意自身的健康状态、生活条件也较为艰苦，同时在疾病初期仅体现为感冒、流涕等情况时未能及时就医，耽误了传染病的预防阶段，故而成为了该传染病的高发人群。

表 24 2004-2016 年内四类人群发病数占当年比重（单位：%）

职业	2004 年	2007 年	2010 年	2013 年	2016 年
学生	6.3464	5.7697	4.9388	4.0953	4.3162
农民	60.1103	60.9406	60.9606	63.8998	63.8141
离退人员	4.9461	4.4498	3.9786	4.1079	4.6588
家务及待业	5.6598	6.0392	7.2919	10.4507	12.7467

同时，根据表 25 所示的“2004-2016 年全国各地区发病率占比（部分）”来看，贵州省、新疆地区的发病率一直居高不下，而西藏、青海则呈上升趋势，这些地区可能由于人口分布过于松散、省内医务人员人数不足、基础医疗设施不完善等问题，同时，社会经济因素也称为影响地区发病率的原因之一。导致了这些地区成为了该传染病的高发区。

表 25 2004-2016 年全国各地区发病率占比（部分）

地区	2004 年	2007 年	2010 年	2013 年	2016 年
贵州省	4.96	6.23	5.89	6.06	6.28
西藏	2.91	3.12	4.78	6.19	7.32
青海省	2.98	3.11	3.61	4.75	6.19
新疆	5.57	6.93	6.79	7.73	8.85
广东省	2.74	3.50	3.84	3.30	3.43

然而广东省作为我国经济第一大省，在该传染病的发病率上也有较高的数据，根据广东省 GDP、居民人均可支配收入、政府地方财政医疗支出、城镇化、流动

人口等数据，结合相关文献材料进行分析，得出导致这种情况的较大原因有两点：1.广东省城镇化水平不高，仍有较多农民从事农业及相关生产；2.广东省流动人口数量较大，且所享有的社会福利待遇不高。这两点原因导致广东省仍维持着较高的发病率。下表为2004-2016年间，贵州、西藏、青海、新疆、广东五省地方财政医疗卫生支出统计表（单位：万元/10万人）。

表 26 2004-2016五省地方财政医疗卫生支出（单位：万元/10万人）

省份	2004年	2007年	2010年	2013年	2016年
贵州省	500.7018	1343.3177	3669.95688	6530.8395	11041.0689
西藏	2308.9492	5938.5121	10680.2	12913.5256	21138.9728
青海省	1176.2152	3533.4420	6916.1811	11875	17379.4266
新疆	992.9648	2186.8926	4739.5835	7107.2968	10693.4945
广东省	800.2403	1457.2391	2911.9998	5348.7786	10199.3817

由上表可以发现，西藏地区政府财政医疗卫生支出居然是最高的，但西藏的发病率不仅不降低，反而显著增长，这个数据非常值得思考。探究具体原因，可能有以下几点：1.该医疗卫生支出可能用于基础医疗设施建设、人才培养、居民的大病治疗补贴等；2.西藏地区经济发展水平较低，省内专业的医务人员人数不足，且主要集中于省内几个较大城市内；3.人口稀少，且分布过于松散，十分不利于医务人员上门诊断或者前往专业的医院进行诊断；4.当地人民对于该传染病的预防、前期症状不了解、不重视。

基于以上分析，现向卫生健康委员会有关部门写一封公开信，谈谈本人对该传染病疫情防治的看法和建议，公开信内容如下：

致卫生健康委员会的一封信

尊敬的卫生健康委员会：

本文通过分析2004年-2016年该传染病的流行病学变化趋势、结合地区、职业、地区经济发展等相关因素，对该传染病的传播特征、未来预测进行了一定的工作，现谈谈对于该传染病疫情防治的看法：

首先，该传染病主要爆发人群为缺乏锻炼、身体素质较弱的学生和离退人员，以及不很在意个人身体、就医意愿较弱的农民、家务及待业等人群。其次，该传染病在经济欠发达、人口分布较为松散、流动人口较多的地区易高发。

针对以上问题，下面给出个人的一些建议：

1. 提高公众对于该传染病的重视程度，在不明显提高地方政府医疗卫生支出的前提下，多做宣传和教育活动，让大众了解该传染病的传播途径、前期症状等，教授该传染病的预防、简易初期诊断方法等。

2. 提高基层医务人员的专业水平和福利待遇，进行一些常见的传染病初步诊断

培训，提高医务人员的业务水平。鼓励经济欠发达地区的医务人员经常前往一些较为偏僻但人口较为聚集的地方，进行集中的传染病诊断等。

3. 向地方政府建议，从财政医疗卫生支出中抽出一部分作为该传染病的专项资金，虽然该传染病死亡率较低，但发病率较高，且伴随较强的传染性和危害性，同时针对的人群又有一定的社会敏感性，建议在预防资金上产生一定偏向。

4. 呼吁大众积极参与身体锻炼和定期体检。积极参与身体锻炼可增强抵抗力、提高自身的身体素质。定期体检可及时发现可能存在的一些隐患，便于医务人员排查一系列传染病，也为群众的身体健康提供保障。

5. 对于重点区域和重点人群，可建立相应的预测模型进行跟踪排查，并根据主要发病原因进行相应的对策处理及疫情预防。

以上是个人对该传染病疫情防治的一些看法和建议。

5 模型的检验与改进

本文通过使用双模型并行预测的方法，利用文中数据使用 BP 神经网络对每年总发病、死亡数进行预测，使用灰色马尔可夫模型对 A、B、C、D 四种方案诊断病例数进行预测。通过比较两个模型的预测结果和误差，得到了 2019 年全国感染该疾病的发病人数和死亡人数。由于使用了双模型并行预测，本文的结果较好的反映了真实的情况。接着，建立了基于自回归求和移动平均（ARMIA）的该传染病传播模型；并通过该模型预测 2019 该传染病排名前 3 位的重点区域和重点人群。然后，本文结合上海地区经济发展的相关公开数据，选择“流动人口、人均可支配收入、医疗支出、城镇人口”四项指标并提出 “可获得医疗帮助的能力”这一新的角度，以此建立多元线性回归模型。并通过对比广东省的数据验证了该模型的可靠性。

针对于模型分析的改进：

针对问题一，我们在建立模型的过程中仅选取了年份作为模型的初始值预测 2019 年全国感染该疾病的发病人数和死亡人数。这里默认了没有其他相关的因素会影响发病人数和死亡人数，下一步的改进是在互联网中寻找其他相关联的数据建立模型，从而得到更好的结果。

针对问题二，建立基于自回归求和移动平均（ARMIA）的该传染病传播模型预测 2019 该传染病防控的区域和职业人群，然而由于本文所给定的数据量较少，缺乏季节性的数据，无法在建模过程中使用较为复杂的模型。下一步的改进通过寻找数据，扩充数据量，从而使用 SARIMA-GRNN 等较复杂的模型进行建模，从而得到更加准确的结果。

针对问题三，我们提出“可获得医疗帮助的能力”这一新的角度，利用相关经济指标建立了多元线性回归模型。然而由于时间仓促，本文仅选取了四个相关指标进行建模，下一步的改进是通过寻找相关的数据，建立更加完善、可靠的模型。

由于时间仓促，本文在模型建立上还存在一些问题，比如，在本文中，数据库中部分数据没有用到，如各年中各诊断方法的数据等，这为我们未来对该传染病各诊疗方案的深入分析理解提供了新的思路。

6 模型的评价与推广

6.1 模型的评价

6.1.1 模型的优点

1. 问题一针对 2004-2016 连续年份的总发病数和总死亡数，采用双模型并行预测——使用 BP 神经网络对每年总发病、死亡数进行预测；使用灰色马尔可夫模型对 A、B、C、D 四种方案诊断病例数进行预测，通过对两个模型预测结果进行效果比较，保证了最终预测结果具有较高准确性。

2.问题二应用地理信息软件 Arcgis10.4 对不同年份下不同地区和不同职业的发病率进行全部和局部空间自相关分析，探讨该传染病的时空聚集特征，通过地图的方式展示非常清晰直观。

3.问题三选取多个经济发展相关指标进行传染病分析，并提出“可获得医疗帮助的能力”这一新的角度，探讨了经济发展对该传染病发病人数与死亡人数的影响，有助于通过改善经济发展能力减少患病率。

4.问题四公开信部分，针对各地区经济发展状况与传染病防控之间的潜在关系，对传染病疫情的防治提出建议，为相关地区制定传染病防控方针提供参考。

6.1.2 模型的缺点

1.本文只选取了“常住人口数、人均可支配收入、医疗支出、城镇化率、国内生产总值（GDP）、流动人口、发病数”7 个传染病与地区经济发展相关指标数据进行分析，还有许多经济发展指标未能考虑，使得统计分析结果不够全面，与真实情况可能存在一定误差。

2.由于时间和专业等原因，搜集的资料可能不够全面，对问题的理解分析以及模型建立上会有所欠缺。

6.2 模型的推广

传染病发病的预测是传染病防治工作中非常重要的一个环节，根据某一传染病变化规律建立的该传染病疫情发展预测模型，有效预测传染病的发病人数及发病率，对该传染病的防治有着重大意义。本文分析了该传染病的流行病学变化趋势，并通过 BP 神经网络模型较为准确的预测未来全国感染该疾病的发病人数和死亡人数，为各地区针对自身患病情况制定防控方案；基于 ARMIA 传染病传播模型预测 2019 年防控重点区域和人群，重点区域以及患传染病的高危职业人群应当引起广泛注意，做好预报保护措施，本文的预测模型可广泛应用于政府及医疗保障机构采取措施提前预警及防控流行性传染病对居民健康的影响，控制并逐渐消灭该传染病，提高全国人民的生活及健康水平。

参考文献

[1]尤爱国,杨建华,赵晓静,等.河南省 2013 年学校传染病突发公共卫生事件流行病学特征[J].郑州大学学报(医学版),2015,50(03):347-350.

[2]祝丙华,王立贵,孙岩松,等.基于大数据传染病监测预警研究进展[J].中国公共卫生,2016,32(09):1276-1279.

[3]曾子明,黄城莺.基于 BP 神经网络的突发传染病舆情热度趋势预测模型研究[J].现代情报,2018,38(05):37-44+52.

[4]郭中凯,任秋艳,李建生.具有年龄结构的 SIR 传染病模型的最优接种和治疗策略[J].南京师大学报(自然科学版),2019,42(01):28-35.

[5]蔡晓虹,万秋萍,吴益生,等.ARIMA 模型预测上海市闸北区手足口病发病趋势[J].实用预防医学,2012,19(3):381-384.

[6]吴莹,刘文东,梁祁,等.江苏省乙型肝炎流行趋势的时间序列分析及预测[J].江苏预防医学,2010,21(6):15-17.

附录

灰色马尔可夫模型代码.m

```
clear,clc;
%A=xlsread('分省年度数据.xls','$K$5:$B$5');
%A=sort(A);
A = [134.60        195.69   162.48   170.35   182.72 ];
syms a b;
c=[a b]';
B=cumsum(A); %  原始数据累加
n=length(A);
for i=1:(n-1)
C(i)=(B(i)+B(i+1))/2; %  生成累加矩阵
end
%  计算待定参数的值
D=A;D(1)=[];
D=D';
E=[-C;ones(1,n-1)];
c=inv(E*E')*E*D;
c=c';
a=c(1);b=c(2);
%  预测后续数据
F=[];F(1)=A(1);
for i=2:(n+4)
F(i)=(A(1)-b/a)/exp(a*(i-1))+b/a;
end
G=[];G(1)=A(1);
for i=2:(n+1)
   G(i)=F(i)-F(i-1); %得到预测出来的数据
end
t1=2004:2008;
t2=2009:2010;
G;a;b;%  输出预测值，发展系数和灰色作用量
disp(G);
scatter(t1,A,'b');
hold on
plot(2004:2010,G,'r');
xlabel('年份');ylabel('人口数/万人');
title('基于灰色预测模型的未来十年湖南省人口趋势图');
text(2026,7444.6,'7444.6 万');
legend('实际人口数量','预测人口数量');
grid on
```

绘制折线图.m

```
x=2004:3:2016;%x 轴上的数据，第一个值代表数据开始，第二个值代表间隔，第三个值代表终止
a=[66.15 98.72 91.86 72.82 70.84 ]; %a 数据 y 值
b=[66.71 98.38 91.22 72.33 70.27 ]; %b 数据 y 值
plot(x,a,'-*b');
hold on;
plot(x,b,'-or');
hold on;
%plot(x,a,'-*b',x,b,'-or'); %线性，颜色，标记
axis([2004,2016,60,100])  %确定 x 轴与 y 轴框图大小
set(gca,'XTick',[2004:3:2016]) %x 轴范围 1-6，间隔 1
set(gca,'YTick',[60:10:100]) %y 轴范围 0-700，间隔 100
legend('真实值','回归值');    %右上角标注
xlabel('年份')  %x 轴坐标描述
ylabel('发病数（/10 万）') %y 轴坐标描述
%2004-2016 年上海地区传染病发病数回归模型示意图
```