2020 年第十三届“认证杯”数学中国
数学建模网络挑战赛第一阶段论文

题　　目　流行病级别划分、传播预测及措施制定

摘　要

新型冠状病毒肺炎 2019-nCoV 给中国乃至全世界都带来了深重的灾难，对世界经济也造成了不可逆的影响。该病毒传染性强、危害较大，需要我们高度警惕。国内目前疫情基本得到控制，但是为避免无症状感染者导致疫情反扑，我们有必要利用相关数学算法，结合大数据背景，开展相关分析，并提出应对措施。

首先我们对附件给出的数据进行分类以及预处理，包括数据的完整性、冗杂性以及相关性分析，这是后续建模的基础。

针对问题一，我们选取了 16 种较为著名的流行病，并考虑了 14 种评价指标。首先，我们选用了 **R 型聚类法**对指标进行降维处理；接着，借助流行病学对传染病的分类，选用了 **Q 型聚类法**将 16 种疾病分成了四大类：散发，暴发、流行和大流行，实现各级别之间的定量化识别。最后，我们根据**主成分分析法**对不同类别传染病进行综合评价，合理量化了“流行”和“大流行”病的界限。

针对问题二，根据问题一中的聚类分析结果以及建立的主成分评价模型对不同地区的感染程度进行分级，然后根据分级结果给出相应的抽检对策。在对无症状感染者进行预测时，以湖北省为例，分别从统计学和流行病学两个方面展开分析。统计学方法，我们选用了**响应面预测模型**，得到了四种因素对无症状感染数的敏感性强弱满足：基本传染数 R_0＞治愈率 R_c＞患病人数 P＞潜伏期 T；流行病学方面，选用了**修正 SEIR 模型**，预测结果表明：继续执行管控措施，大约到 5 月中下旬，日新增无症状感染者人数降为 0。该结论与响应面预测结果基本一致，说明了预测模型合理，预测结果准确性较高。

最后，我们向世卫组织写了一封建议信，阐述了我们对该病毒的认识并给出了相关对策。

关 键 词： 聚类分析；主成分分析；响应面预测；修正 SEIR 模型；2019-nCoV

Abstract

The new coronavirus pneumonia 2019-nCoV has brought profound disasters to China and the whole world, and has also caused an irreversible impact on the world economy. The virus is highly contagious and harmful, and requires us to be highly vigilant. The current epidemic situation in China is basically under control, but in order to avoid asymptomatic infections causing the epidemic to counterattack, it is necessary for us to use relevant mathematical algorithms, combined with the background of big data, to carry out relevant analysis and propose countermeasures.

First, we classify and preprocess the data given in the attachment, including data integrity, redundancy, and correlation analysis, which is the basis of subsequent modeling.

In response to question 1, we selected 16 relatively famous epidemics and considered 14 evaluation indicators. First, we selected the R-type clustering method to reduce the dimensionality of the indicators; then, based on the epidemiological classification of the epidemic, the Q-type clustering method was used to divide the 16 diseases into four categories: sporadic, outbreak, popularity and pandemic, to achieve quantitative identification between all levels. Finally, we conducted a comprehensive evaluation of different types of epidemics according to the principal component analysis method, and reasonably quantified the boundaries between "epidemic" and "pandemic" diseases.

For problem two, according to the cluster analysis results in problem one and the established principal component evaluation model, the infection degree in different regions is graded, and then the corresponding sampling countermeasures are given according to the classification results. In the prediction of asymptomatic infections, taking Hubei Province as an example, the analysis is carried out from two aspects of statistics and epidemiology. Statistical methods, we chose the response surface prediction model, and obtained the sensitivity of four factors to the number of asymptomatic infections: basic infection number R0> cure rate Rc> number of patients P> latent period T; epidemiology The revised SEIR model was selected, and the prediction results indicated that the control measures will continue to be implemented, and the number of newly-increased asymptomatic infections will drop to zero around mid to late May. This is basically consistent with the previous response surface prediction results, indicating that the prediction model is reasonable and the prediction results are of higher accuracy.

Finally, we wrote a letter of recommendation to WHO, explaining our understanding of the virus and giving relevant countermeasures.

Key words: Cluster analysis; Principal component analysis; Response surface prediction; Modified SEIR model; 2019-nCoV

目　　录

一、问题重述 2
1.1 问题背景 2
1.2 问题分析 2
二、模型假设 2
三、符号说明 3
四、技术路线 3
五、数据分析与预处理 4
六、问题一模型建立与求解 6
6.1 问题一分析 6
6.2 聚类模型建立与求解 7
6.2.1 R 型聚类模型建立 7
6.2.2 R 型聚类模型求解与分析 9
6.2.3 Q 型聚类模型建立 10
6.2.4 Q 型聚类模型求解与分析 11
6.3 综合评价模型建立与求解 12
6.3.1 综合评价模型建立 12
6.3.2 综合评价模型求解与分析 14
6.4 问题一小结 15
七、问题二模型建立与求解 15
7.1 问题二分析 15
7.2 病毒抽检对策 16
7.3 预测模型建立与求解 16
7.3.1 响应面预测模型建立 16
7.3.2 响应面预测模型求解与分析 17
7.3.3 修正 SEIR 预测模型建立 21
7.3.3 修正 SEIR 预测模型求解与分析 22
7.4 问题二小结 23
7.5 模型敏感性分析 23
八、给世卫组织(WHO)的一封信 24
九、模型评价与推广 25
9.1 模型评价 25
9.2 模型改进 25
参考文献 26
附录 27

一、问题重述

1.1 问题背景

新型冠状病毒肺炎 2019-nCoV 给中国乃至全世界都带来了深重的灾难，对世界经济也造成了不可逆的影响。该病毒传染性强、危害较大，且目前并没有特效药治疗，需要我们高度警惕。在中国政府、全国人民的积极应对下，中国国内疫情已得到基本控制，但是全球整体情况却不容乐观。目前，在全球已有超过 200 个国家/地区报告了病毒感染病例，累计确诊超过 200 万，这一数字值得我们重视。

2019-nCoV 病毒潜伏期较长，且存在无症状感染者，即指无临床症状、但呼吸道等标本新冠病毒病原学检测呈阳性者。相关报道指出，一名感染者从未出现症状，但所释放的病毒量与出现症状的人相当。因此，一部分科学家猜测：一些感染者“在症状轻微或无症状时具有高度传染性”。张文宏团队预测，以目前部分研究为例，感染新冠病毒的人群中，无症状感染者的比例大约为 18%~31%，这一结论告诉我们需要重视无症状感染者这一患病群体。因此，为避免无症状感染者导致疫情反扑，我们有必要利用相关数学算法，结合大数据背景，分析预测无症状感染者未来发展趋势，从而提出针对性的应对措施[1][2]。

1.2 问题分析

问题一要求我们给出合理的界定“流行” (Epidemic) 和“大流行”(Pandemic) 病的定量条件。由于目前针对流行病分级还是以定性分析为主，因此，为定量分析“流行”与“大流行”，我们需要对传统的分级进行一定的数学处理。通过查阅相关资料，在流行病学上，疾病的传播强度可以简单分为四级，分别是散发（sporadic），暴发（outbreak）、流行（epidemic）和大流行（pandemic）[3-7]。这一分类对流行病进行了定量的划分，基于此类别，我们可参考历年重大流行病的传播强度，以若干评价指标展开聚类分析，将流行病分成对应的 4 类。然后再根据分类结果，利用主成分分析，对各个类别展开综合打分。最后，根据打分结果，获取各级别流行病所在区间，从而定量的区分“流行”(Epidemic)和“大流行”(Pandemic)。

问题二要求我们针对某地区，给出切实可行的病毒检测抽样方案，并给出无症状感染者分布预测模型和针对相应预测结果的应对方案。我们可借助问题一中聚类分析提取的指标，利用主成分评价模型对该地区患病严重程度进行综合打分，根据打分结果的高低制定具有针对性的监测和隔离措施。为实现对无症状感染者的分布与预测，可考虑两种方法，第一种是统计学方法：首先，我们以无症状感染人数为因变量，根据 R 型聚类结果选取 4 种典型指标（这里选取患病人数、基本传染数、潜伏期、治愈率四个指标），建立因变量与自变量的数学表达式。第二种方法选择考虑了病毒潜伏期的传播模型 SEIR，对后期无症状感染者进行预测。最后对比分析两种方法的预测结果，并给出应对方案。

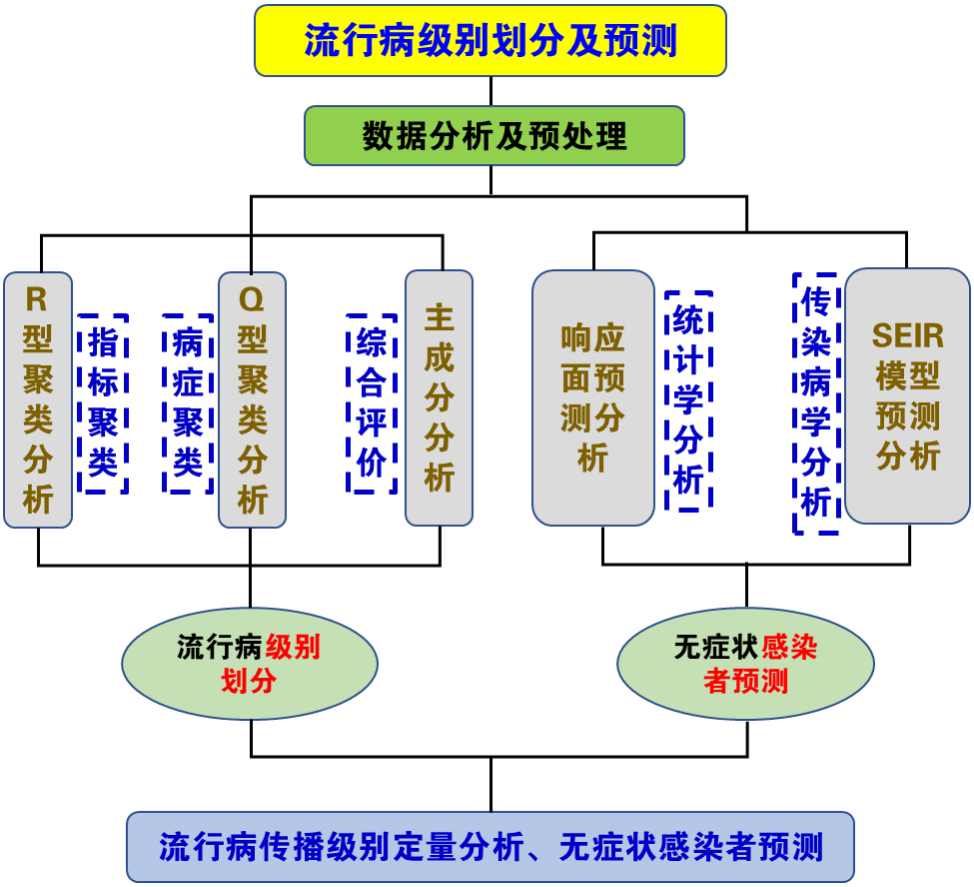
二、模型假设

1. 假设题目所给的数据真实可靠；
2. 假设短期内无特效药出现；
3. 假设短期内不存在大量的输入病例。

三、符号说明

符号	符号说明	备注
C_i	第 i 个变量数据完整率	%
P	暴发地综合人口数	万人
I	感染人数	人
D	病死人数	人
R	治愈人数	人
T	疫情持续时间	天
N	感染国家数量	个
r	增长率	%
R_0	基本传染数	/
R_d	病死率	%
D_p	暴发地人口密度	人/平方千米
Z	主成分分析模型综合得分值	/
t	潜伏期时间	天
R_w	无症状感染者比例	%

四、技术路线



五、数据分析与预处理

由于附件给出的数据量非常庞大冗杂，为使得建立的分析模型能够更加真实反应传染病的传播能力强弱，在对问题进行正式求解之前，先对数据进行分析，并通过相应的数学方法进行数据预处理，这有助于后续模型的建立与求解。

通过对数据体分析，将变量主要分为三类：文本说明变量，定量描述变量，事件分类变量[8]。

$$\text{变量类型}=\begin{cases}\text{文本说明变量}\\ \text{定量描述变量}\\ \text{事件分类变量}\end{cases}$$

文本说明变量主要指数据的基本特征和完整情况，如在附件给定的数据中时间（Date）、地点（Country）等；定量描述变量主要描述数据的真实值，如 Afghanistan 的国家人口总数（Population）为 3900 万、全世界 2020 年 1 月 22 日确诊人数（Confirmed）为 555 人等；事件分类变量指数据的类别，如在附件 reference 中，UID 为 4 时，对应的 iso 为 AF、AFG 等。

1. 数据完整性分析

本此建模统计了各个变量的数据完整性，用数据完整率C_{i}表示，即第 i 个变量的数据完整性。

$$C_{\mathrm{i}}=\frac{\text{第i个指标有效变量之和}}{\text{第i指标变量数总和}}$$

根据此定义可知，C_{i}越小，变量 i 数据的完整性越高，越具有真实性，可研究性也越强。在本次建模中，附件提供的相关参数主要包括：综合人口数（P）、感染人数（I）、病死人数（D）、治愈人数（R）、疫情持续时间（T）、感染国家数量（N）、增长率（r）等。这里统计了附件给出的部分指标的数据完整性，统计结果如表 5.1 所示。

表 5.1 指标完整率统计结果

指标	综合人口数（P）	感染人数（I）	病死人数（D）	治愈人数（R）	持续时间（T）	感染国家数（N）	增长率（r）
完整率	1.00	1.00	0.99	0.99	1.00	1.00	0.99

由表 5.1 可知，附件给出的参数数据完整率C_{i}＞99%，甚至为 1（如综合人口数（P））。因此在进行问题分析时，应充分利用这些数据。

2. “冗余变量”分析

由于附件给定数据较多，有必要将“冗余变量”进行降维处理，并对无效变量进行剔除。在给定附件中 reference 中，iso2、iso3 以及 code3 数据可暂时不予考虑，因此，建模时可剔除该因素。

3. 变量相关性分析

为避免变量间存在重复的可能性，对以上变量进行关联程度分析。用相似系数来衡量变量之间的相似程度（关联度），若用$C_{\alpha,\beta}$表示变量之间的相似系数，则应该满足：

$$\left|C_{\alpha,\beta}\right|\le 1,\ \text{且}\ C_{\alpha,\alpha}=1$$

$$C_{\alpha,\beta}=1,\ \text{当且仅当}\ \alpha=k\beta,\ k\neq 0$$

$$C_{\alpha,\beta}=C_{\beta,\alpha}$$

（1）夹角余弦相关度分析

余弦相似度是通过计算两个向量之间的夹角余弦值来评估其相似度。若余弦值越接近 1，则向量之间的夹角越小，他们的方向更加一致。设有向量 $\vec{a}$、$\vec{b}$，则其夹角余弦 $\cos\theta$：

$$\cos\theta=\frac{\vec{a}\cdot\vec{b}}{\left\|\vec{a}\right\|\left\|\vec{b}\right\|} \tag{5.1}$$

代入坐标 (x_1,y_1)、(x_2,y_2) 得：

$$\cos\theta=\frac{x_1x_2+y_1y_2}{\sqrt{x_1^2+y_1^2}\bullet\sqrt{x_2^2+y_2^2}} \tag{5.2}$$

设向量 $\vec{A}=(A_1,A_2,\cdots,A_n)$，$\vec{B}=(B_1,B_2,\cdots,B_n)$，则夹角余弦：

$$\cos\theta=\frac{\sum_{i=1}^{n}(A_i\times B_i)}{\sqrt{\sum_{i=1}^{n}A_i^2}\times\sqrt{\sum_{i=1}^{n}B_i^2}} \tag{5.3}$$

（2）皮尔逊相关系数分析

皮尔逊相关系数定义为两个变量之间的协方差和标准差的商，设变量 P_i、P_j 之间的相关系数为 $R_{i,j}$，其表达式如下：

$$\rho_{i,j}=\frac{\operatorname{cov}(P_i,P_j)}{\sigma_{p_i}\sigma_{p_j}}=\frac{E\left[(P_i-\mu P_i)(P_j-\mu P_j)\right]}{\sigma_{P_i}\sigma_{P_j}} \tag{5.4}$$

估算样本的协方差和标准差，可以得到皮尔逊相关系数 $R_{i,j}$：

$$\begin{aligned}R_{i,j}&=\frac{(P_i-\bar{P}_i,P_j-\bar{P}_j)}{\left\|P_i-\bar{P}_j\right\|\left\|P_j-\bar{P}_j\right\|}\\&=\frac{\sum_{m=1}^{n}(P_{im}-\bar{P}_i)(P_{jm}-\bar{P}_j)}{\left[\sum_{m=1}^{n}(P_{im}-\bar{P}_i)^2\cdot\sum_{m=1}^{n}(P_{jm}-\bar{P}_j)^2\right]^{\frac{1}{2}}},\quad -1\le R_{i,j}\le 1\end{aligned} \tag{5.5}$$

由于夹角余弦去中心化后，得到皮尔逊相关系数，考虑本题数据受级别膨胀影响（某些变量数据之间存在数量级的差异）故采用皮尔逊相关分析所得结果更具有可接受性。通过编程，部分变量之间相关性结果如表 5.2 所示。

表 5.2 变量皮尔逊相关系数分析

指标	P	I	D	R	T	N	r
P	1	0.6566	0.6187	0.2134	0.6375	0.5154	0.6002
I	0.6566	1	0.8368	-0.0740	0.8648	0.5080	0.7057
D	0.6187	0.8368	1	-0.7465	0.7051	0.6273	0.7215
R	0.2134	-0.0740	-0.7465	1	0.1711	-0.4148	-0.4133
T	0.6375	0.8648	0.7051	0.1711	1	0.9362	0.8388
N	0.5154	0.5080	0.6273	-0.4148	0.9362	1	0.8912
r	0.6002	0.7057	0.7215	-0.4133	0.8388	0.8912	1

根据得到的相关系数表（表 5.2），可以看出，感染人数（I）越多，病死人数（D）越多，对应的持续时间越长（T）；疫情持续时间越长，对应的感染国家数量（N）越多；感染人数（I）越多，对应的增长率（r）越高；综合人口数与感染人数呈正相关，但是相关性不强，相关系数仅为 0.66。这些结论基本符合常规认识，也说明了传染病传播能力强弱与多个因素有关，在进行后续分析时，应考虑多个因素的影响。

六、问题一模型建立与求解

6.1 问题一分析

问题一要求我们给出合理的界定“流行” (Epidemic) 和“大流行”(Pandemic) 病的定量条件。由于目前针对流行病分级还是以定性分析为主，因此，为定量分析“流行”与“大流行”，我们需要对传统的分级进行一定的数学处理。

在流行病学上，疾病的传播强度可以简单分为四级，分别是散发（sporadic），暴发（outbreak）、流行（epidemic）和大流行（pandemic）。其中“散发”指病例之间没有时间或空间上的关联；“暴发”指一个集体单位或局部地区中，短时间内出现了很多同样的病人；如果某疾病的发病率较通常水平显著升高，就可以认为疾病处于“流行”状态；而当疾病迅速蔓延，短时间内跨越国界甚至洲界时，就可以称之为“大流行”。

这一分类对流行病进行了定量的划分，基于此类别，我们可参考历年重大流行病的传播强度，以若干评价指标展开聚类分析，分成对应的 4 类。然后再根据分类结果，利用主成分分析，对各个类别展开综合打分。最后，根据打分分值，获取各级别流行病所在区间，从而定量的区分“流行”(Epidemic)和“大流行”(Pandemic)。本问题基本流程图如图 6.1 所示。

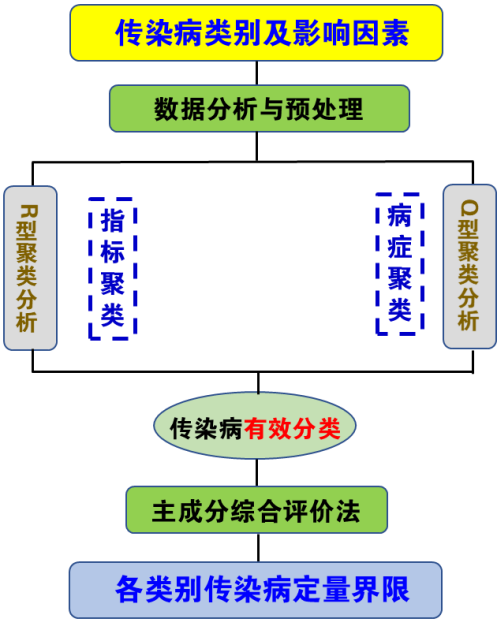


图 6.1 问题一求解思路

6.2 聚类模型建立与求解

聚类分析法又称群分析，是对多个样本或指标进行定量分类的一种多元统计分析方法[9]。该方法可用 R 行聚类分析对指标进行分类，再用 Q 型聚类分析对样本进行分类，从而帮助我们识别流行病的传播级别。

由于评价指标种类繁多，单一传染病对应的指标冗杂，我们先选用 R 型聚类法对评价指标进行分析，基于该结果，再利用 Q 型聚类法对传染病级别进行分类，并识别其特征。

6.2.1 R 型聚类模型建立

R 型聚类分析主要用于对指标进行降维处理。因此，我们首先需要选择影响病毒传播能力的指标。通过文献调研以及赛题的背景说明，我们主要考虑 16 项较为熟知的传染病（具体见表 6.1），每种传染病考虑 14 项评价指标，分别为：暴发地综合人口数（P）、感染人数（I）、病死人数（D）、治愈人数（R）、疫情持续时间（T）、感染国家数量（N）、增长率（r）、基本传染数（R_0）、病死率（R_d）、暴发地人均 GDP、暴发地人口密度（D_p）、暴发地防疫措施（M）。潜伏期时间（t）、无症状感染者比例（R_w）。

特别值得注意的是，这里引入了一个流行病学方面的指标：基本传染数（R_0），又名（Basic reproduction number）。该指标表示在没有外力介入，所有人都没有免疫力的情况下，一个感染某种传染病的人，会传染给其他多少个人的平均数。R_0 的一个重要临界点是 R0=1，R_0 的数字越大，代表流行病的越难控制。如当 R_0 小于 1 时，表示传染病将会逐渐消失；当 R_0 等于 1 时，传染病会变成的地方性流行病；当 R_0 大于 1 时，传染病以指数方式散布，但是不会永远持续，因为要么所有人均被感染，或者是有些人病愈后产生了免疫力[7]，常见传染病的传播途径及 R_0 值见表 6.1。

表 6.1 16 种常见传染病传播途径及 R_0 值

序号	传染病名称	传播途径	基本传染数（R_0）
1	乙肝	血液传播、性传播	1-2
2	鼠疫	皮肤传播、飞沫传播	1-3
3	霍乱	消化道传播	1-2
4	麻疹	空气传播	12-18
5	白喉	唾液	6-7
6	天花	空气传播、飞沫传播	5-7
7	脊髓灰质炎	粪口传播	5-7
8	风疹	空气传播、飞沫传播	5-7
9	流行性腮腺炎	空气传播、飞沫传播	4-7
10	HIV/AIDS	性传播	2-5
11	百日咳	空气传播、飞沫传播	5.5
12	SARS	空气传播、飞沫传播	0.85-3
13	流行性感冒	空气传播、飞沫传播	2-3
14	MERS	空气传播、飞沫传播	1-2
15	HIN1	空气传播、飞沫传播	1.75
16	COVID-19	空气传播、飞沫传播	3.77

（1）变量相似性分析

在对变量进行聚类分析时，首先要确定变量的相似性度量，常用的变量相似性度量有两种，分别是相关系数方法和夹角余弦法。由于在对变量进行聚类分析时，利用相关系数矩阵较多，为此，本次建模选用相关系数法。该方法介绍如下：

记变量 $\mathbf{x}_j$ 的取值（x_{ij}，x_{2j}，…，x_{nj}）$^T \in \mathbf{R}^n$（j=1，2，...，m）。则可以用两变量 $\mathbf{x}_j$ 与 $\mathbf{x}_k$ 的样本相关系数作为它们的相似性度量，即：

$$r_{jk}=\frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)}{[\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2]^{\frac{1}{2}}} \quad (6.1)$$

（2）变量聚类法

类似于样本集合聚类分析中最常用的最短距离法、最长距离法等，变量聚类法采用了与系统聚类法相同的思路和过程。在变量聚类问题中，常用的有最长距离法、最短距离法等。

① 最长距离法。

在最长距离法中，定义两类变量的距离为：

$$R(G_1，G_2)=\max\{d_{jk}\}，x_j\in G_1，x_k\in G_2 \quad (6.2)$$

式中，$d_{jk}=1-|r_{jk}|$或 $d_{jk}^2=1-r_{jk}^2$，这时，$R(G_1，G_2)$与两类中相似性最小的两个变量间的相似度量值有关。

② 最短距离法

在最长距离法中，定义两类变量的距离为：

$$R(G_1，G_2)=\min\{d_{jk}\}，x_j\in G_1，x_k\in G_2 \quad (6.3)$$

式中，$d_{jk}=1-|r_{jk}|$或 $d_{jk}^2=1-r_{jk}^2$，这时，$R(G_1，G_2)$与两类中相似性最大的两个变量间的相似度量值有关。

该部分流程图如图 6.2 所示。

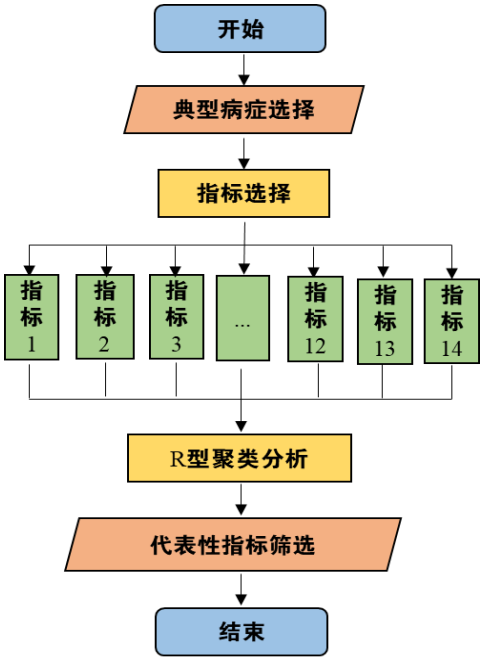


图 6.2 R 型聚类分析流程图

6.2.2 R 型聚类模型求解与分析

通过中国疾控预防控制中心提供的数据，对选择的 14 项指标进行 R 型聚类分析，再选择每个指标中的代表性指标。首先对每个指标的数据分别进行标准化处理。变量间的相似性度量的计算选用类平均法。

通过 MATLAB 编程，得到的聚类树型图如图 6.3 所示。

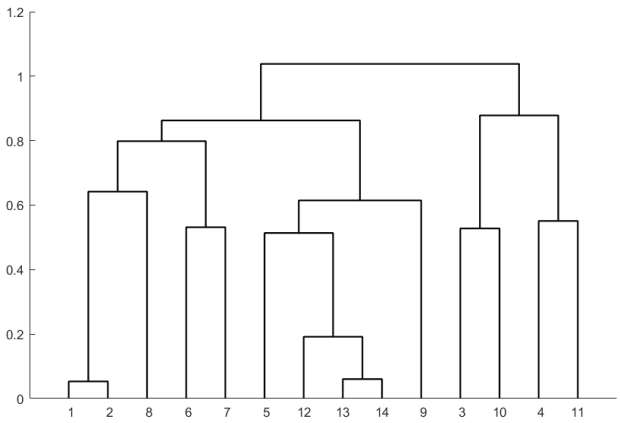


图 6.3 指标聚类树形图

从图 6.3 可以看出，指标 1、2、8 具有较大的相关性，最先被聚到了一起。将 18 个指标分为 7 类，每一类具体指标以及最终所确定的 7 个代表性指标，如表 6.2 所示。

表 6.2 各类别指标基本信息

序号	指标	代表性指标
第 1 类	4、11	4（治愈人数）
第 2 类	3、10	3（死亡人数）
第 3 类	9	9（病死率）
第 4 类	5、12、13、14	13（无症状潜伏时间）
第 5 类	6、7	6（感染国家数）
第 6 类	8	8（基本传染数）
第 7 类	1、2	2（感染人数）

从以上结果可以看出，指标 8（基本传染数）、9（病死率）与其它类型病症存在区别，各自单独为一类；第四类包含的指标较多 5（持续时间）、12（暴发地防疫措施）、13（无症状潜伏时间）、14（无症状感染比例），说明这几类指标具有一定的内在联系。一般来说，基本传染数和病死率直接影响传染病的危害程度，暴发地防疫措施与传染病持续时间呈正相关。这些认识在分类结果中均有体现，这也说明了本模型对于该问题具有较好的适应性。

6.2.3 Q 型聚类模型建立

根据 6.2.2 部分得出的结论，利用 Q 型聚类分析对 16 类经典传染病进行分类。

该模型介绍如下。

1. 样本的相似性度量

要用数量化的方法对事物进行分类，就必须用数量化的方法描述事物之间的相似程度。一个事物常常需要用多个变量来刻画。如果对于一群有待分类的样本点需用 p 个变量描述，则每个样本点可以看成是 $\mathbf{R}^p$ 空间中的一个点。因此，很自然地想到可以用距离来度量样本点间的相似程度。

记 $\boldsymbol{\Omega}$ 是样本点集，距离 d（·，·）是 $\boldsymbol{\Omega}\times\boldsymbol{\Omega}\to\mathbf{R}^+$的一个函数，满足条件：

（1）d(x,y)≥0，x，y∈$\boldsymbol{\Omega}$

（2）d(x,y)=0 当且仅当 x=y

（3）d(x,y)=d(y,x)，x，y∈$\boldsymbol{\Omega}$

（4）d(x,y)≤d(x,z)+d(z,y)，x，y，z∈$\boldsymbol{\Omega}$

这一距离的定义是我们所熟知的，它满足正定性、对称性和三角不等式。在聚类分析中，对于定量变量，最常用的是闵氏（Minkowski)距离，即

$$d_{ij}(q)=\left(\sum_{k=1}^{p}\left|X_{ik}-X_{jk}\right|^{q}\right)^{1/q} \qquad (6.4)$$

通过变换 Minkowski 距离中的 q 值，可以产生以下几种不同的距离：

（1）绝对距离 $(q=1)$

$$d_{ij}(1)=\sum_{k=1}^{p}\left|X_{ik}-X_{jk}\right| \qquad (6.3)$$

（2）欧式(Euclidian)距离 $(q=2)$

$$d_{ij}(2)=\left(\sum_{k=1}^{p}\left|X_{ik}-X_{jk}\right|^{2}\right)^{1/2} \qquad (6.4)$$

（3）切比雪夫(Chebychev)距离 (∞)

$$d_{ij}(\infty)=\max_{1\le k\le p}\left|X_{ik}-X_{jk}\right| \tag{6.5}$$

欧氏距离是常用的距离，本文在使用 MATLAB 进行系统类聚分析时便选用欧式距离。

2. 类与类之间的相似性度量

如果有两个样本类 G_1 和 G_2，存在多种方法度量它们之间的距离。常用的类间距离定义有 8 种，分别为最近距离法、最远距离法、中间距离法、重心法、类平均法、可变类平均法、可变法和离差平方和法，与之对应的系统聚类方法也有 8 种。这里我们用类间平均连接法进行系统聚类，类间平均连接法定义类间距离平方为两类中元素两两之间距离平方的平均值，即：

$$D_{pq}^{\ 2}=\frac{1}{n_p n_q}\sum_{X_i\in G_p}\sum_{X_j\in G_j}d_{ij}^{\ 2} \tag{6.6}$$

设类G_p和类G_q合并为新类G_r，则任意类G_k和G_r的距离为：

$$\begin{aligned}D_{kr}^{\ 2}&=\frac{1}{n_k n_r}\sum_{X_i\in G_p}\sum_{X_j\in G_j}d_{ij}^{\ 2}\\&=\frac{1}{n_k n_r}\Big(\sum_{X_i\in G_k}\sum_{X_j\in G_p}d_{ij}^{\ 2}+\sum_{X_i\in G_k}\sum_{X_j\in G_q}d_{ij}^{\ 2}\Big)\\&=\frac{n_p}{n_r}D_{kp}^{\ 2}+\frac{n_q}{n_r}D_{kq}^{\ 2}\end{aligned} \tag{6.7}$$

3. 聚类图生成

具体生成步骤如下：

（1）计算 n 个样本点两两之间的距离$\{d_{ij}\}$，记为矩阵 $\boldsymbol{D}=(d_{ij})_{n\times n}$。

（2）首先构造 n 个类，每一个类中只包含一个样本点，每一类的平台高度均为 0。

（3）合并距离最近的两类为新类，并且以这两类间的距离值作为聚类图中的平台高度。

（4）计算新类与当前各类的距离，若类的个数已经等于 1，转入步骤（5），否则回到步骤（3)。

（5）画聚类图。

（6）决定类的个数和类。

6.2.4 Q 型聚类模型求解与分析

根据前面 R 型聚类分析法确定的 7 个指标对 16 类传染病进行聚类分析。同样，首先对每个变量的数据分别进行标准化处理，样本间相似性采用欧氏距离度量，类间距离的计算选用类平均法。

通过 MATLAB 编程，得到的聚类树型图如图 6.4 所示。

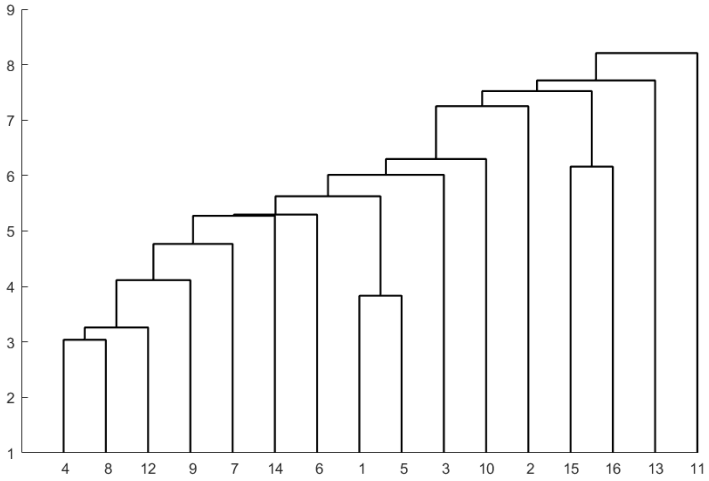


图 6.4 16 种主要传染病聚类树形图

通过 6.4 的聚类结果，我们对不同类别传染病进行统计分析，具体分类结果见表 6.3 所示。

表 6.3 各类别病症基本信息

序号	传染病代号	代表病症	传播级别
第 1 类	11	11（百日咳）	散发（sporadic）
第 2 类	13	13（流行感冒）	暴发（outbreak）
第 3 类	15、16	16（COVID-19）	大流行（epidemic）
第 4 类	4、8、12、9、7、14、6、1、5、3、10、2	4（麻疹）	流行（pandemic）

从以上结果可以看出，传染病 11（百日咳）、13（流行感冒）单独分为一类，传播级别分别为“散发”和“暴发”，传染病 15（甲型 H1N1）、16（COVID-19）分为一类，传播级别为“大流行”，第四类囊括了大部分的传染病（＞50%），传播级别为“流行”。通过查阅相关资料，百日咳是一种急性呼吸道传染病，但是很少能够通过外界条件传染，因此将其归为一类；而流行感冒能引起较大范围传播，但是传播途径易阻断（勤洗手、戴口罩），且病死率低，危害程度较小，归为一类；甲型 H1N1 和 COVID-19 传播能力强，患病致死率较高，传播途径较多，影响程度大，将其归为一类，其分类结果与世卫组织（WHO）给出的结论一致。这符合我们的认识，也说明了本模型具有较高的精度。

6.3 综合评价模型建立与求解

在 6.2 小节中，我们通过聚类分析将 16 类传染病分成了 4 类，并从 14 类指标中筛选出了 7 类影响较大的指标，这部分属于定性分析。接下来，我们根据不同的类别，利用综合评价模型对不同类别的传染病进行打分，实现类别的定量划分。

由于量化指标较多，数据较大，我们选用主成分分析模型作为传染病分类评价模型。评价指标有 2（感染人数）、3（死亡人数）、4（治愈人数）、6（感染国家数）、8（基本传染数）、9（病死率）、13（无症状潜伏时间）。

6.3.1 综合评价模型建立

主成分分析评价指标共有五个，分别用 $x_1, x_2, \cdots, x_7$ 来表示，用 i=1，2，3…,n 分别表

示传染病 i，传染病 i 的指标 $x_1,x_2,\cdots,x_7$ 的取值分别记作 $[a_{i1},a_{i2},a_{i3},a_{i4},a_{i7}]$,构成矩阵 $A=(a_{ij})_{n\times 7}$。

基于主成分分析法的评价模型具体步骤如下[9]：

（1）将统计所得的原始数据进行标准化处理。标准值 $\tilde{a}_{ij}$ 实现过程满足：

$$\tilde{a}_{ij}=\frac{a_{ij}-\mu_j}{s_j},i=1,2,\cdots,n,\ j=1,2,\cdots,7 \tag{6.8}$$

其中：$\mu_j=\frac{1}{n}\sum_{i=1}^{n}a_{ij}$，$s_j=\sqrt{\frac{1}{10-1}\sum_{i=1}^{n}(a_{ij}-\mu_j)^2}$，$j=1,2,\cdots,7$，即 μ_j，s_j 为第 j 个指标的样本均值和样本标准差。对应地，称：

$$\tilde{x}_j=\frac{x_j-\mu_j}{s_j},\ j=1,2,\cdots,7 \tag{6.9}$$

（2）计算相关矩阵 R，其相关系数矩阵 $R=(r_{kj})_{7\times 7}$，满足：

$$r_{kj}=\frac{\sum_{i=1}^{n}\tilde{a}_{ki}\cdot\tilde{a}_{kj}}{n-1},k,\ j=1,2,\cdots,7 \tag{6.10}$$

其中：$r_{kk}=1$，$r_{kj}=r_{jk}$ 是第 k 指标与 j 指标的相关系数。

（3）计算特征值和特征向量。计算相关系数矩阵 R 的特征值，该特征值满足 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_5\geq 0$，同时计算对应的标准化特征向量 μ_1，μ_2,$\cdots$，μ_7，其中 $\mu_j=[\mu_{1j},\ \mu_{2j},\cdots,\ \mu_{7j}]^T$。

特征向量组成 7 个新的指标向量，指标向量满足：

$$Y=[\mathbf{y_1}\ \mathbf{y_2}\ \ldots\ \mathbf{y}_6\ \mathbf{y}_7]^{\mathrm{T}}=\boldsymbol{u}\bullet[x_1'\ x_2'\ \ldots\ x_6'\ x_7']^{\mathrm{T}} \tag{6.11}$$

其中：y_1 是第 1 主成分，y_2 是第 2 主成分，以此类推，y_7 是第 7 主成分。

选择 $p(p\leq 7)$ 个主成分，计算综合评价值。

1）计算特征值 $\lambda_j(j=1,2,\cdots,7)$ 的信息贡献率和累积贡献率。称：

$$b_j=\frac{\lambda_j}{\sum_{k=1}^{7}\lambda_k},\ j=1,2,\cdots,7 \tag{6.12}$$

为主成分 y_j 的信息贡献率；而且称

$$\alpha_p=\frac{\sum_{k=1}^{p}\lambda_k}{\sum_{k=1}^{7}\lambda_k} \tag{6.13}$$

为主成分 y_1，y_2,$\cdots$，y_p 的累积贡献率；并称：

$$\alpha_p=\frac{\sum_{k=1}^{p}\lambda_k}{\sum_{k=1}^{7}\lambda_k} \tag{6.14}$$

2）利用主成分评价模型计算综合得分，并对不同传染病进行打分排序，得到不同类型传染病数学界限。

6.3.2 综合评价模型求解与分析

根据 MATLAB 求解，主成分 y_1，y_2,…，y_p 的累积贡献率具体数据如表 6.4 所示，主成分累计贡献图如图 6.5 所示。

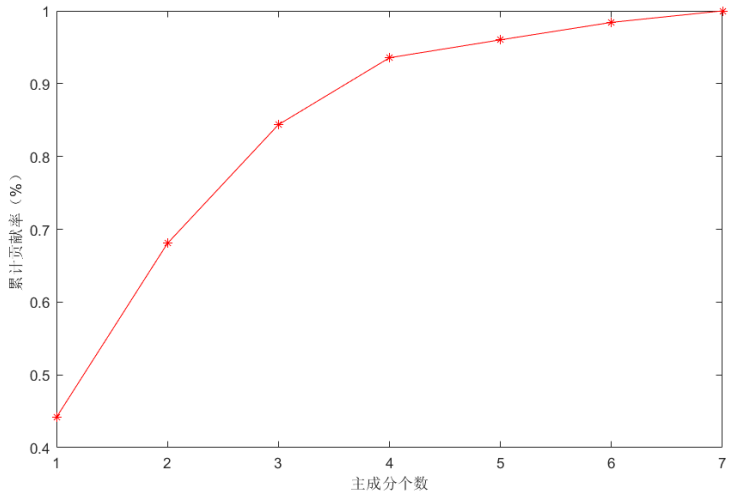
表 6.4 贡献率和累积贡献率结果

序号	特征值	贡献率	累计贡献率	序号	特征值	贡献率	累计贡献率
1	44.1742	0.4417	0.4417	5	2.4643	0.0246	0.9602
2	23.8824	0.2388	0.6806	6	2.2951	0.0240	0.9842
3	16.3155	0.1632	0.8437	7	1.6052	0.0158	1
4	9.1831	0.0918	0.9356				

从表 6.4 知，当 $p \geq 4$ 时，累计贡献率已超过 90%。因此，我们选择前 4 个指标变量 y_1, y_2, y_3, y_4 作为主成分代替 7 个指标变量，进行综合分析。

通过计算，主成分综合评价模型可表示为：

$$Z = 0.4417y_1 + 0.2388y_2 + 0.1632y_3 + 0.0918y_4 \quad (6.15)$$



6.5 主成分累计贡献率分布图

根据对评分结果进行归一化处理，得到了四种类别综合得分，绘制得分分布图如图 6.6 所示。

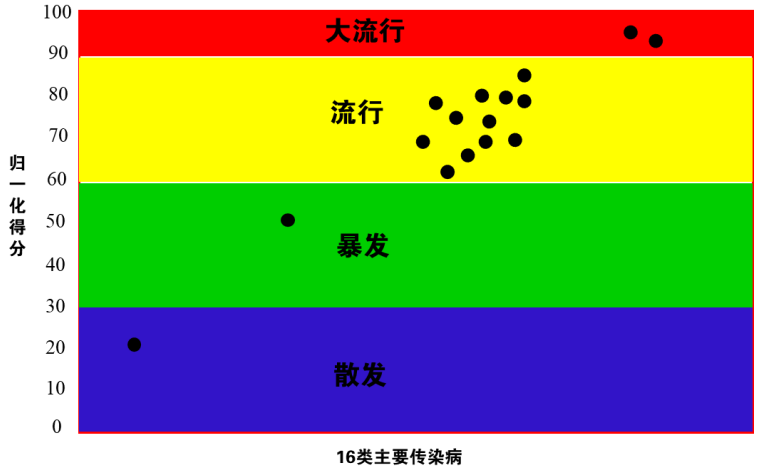


图 6.6 各类传染病综合得分结果分布

根据图 6.6，我们可以看出归一化得分在区间[0，30]时对应的传播类型为“散发”，归一化得分在区间（30，60]时对应的传播类型为“暴发”，归一化得分在区间（60，90]时对应的传播类型为“流行”，归一化得分在区间（90，100]时对应的传播类型为“大流行”。问题一中所述的传染病均为传播较为广泛的疾病，其中甲型 H1N1 和 COVID-19 给人类经济生产带来的影响最大，因此被归纳为“大流行”传播级别。这一分类结果符合我们的认识，具有一定的合理性。

6.4 问题一小结

问题一选择了 16 种较为著名的流行病，并考虑了 14 种评价指标。由于评价指标较多，我们首先选用了 R 型聚类法对 14 种指标进行降维处理；接着，凭借流行病学对流行病的分类，选用了 Q 型聚类法对 16 种疾病进行分类，实现各级别之间的定量化识别。

通过主成分分析对不同级别的流行病进行综合评价，并对打分结果进行归一化处理，得到了各种流行病的分值。我们根据分值，划分了不同传播级别的数学界限，并合理量化了“流行”(Epidemic) 和“大流行”(Pandemic) 病的界限。这对未来流行病传播风险的识别与预判提供了数学基础。

七、问题二模型建立与求解

7.1 问题二分析

问题一通过 R 型聚类筛选出了几类具有代表性的传染病评价指标，这些指标是从世界范围内进行分析的。问题二中要求我们针对一两个国家或者地区展开分析，因此，我们需要对问题一中的指标进行“再提取”。对于病毒监测抽查方案，我们可根据提取的指标，利用主成分评价模型对该地区患病严重程度进行综合打分，根据打分结果的高低制定具有针对性的监测和隔离措施。

为实现对无症状感染者的分布与预测，首先，我们以无症状感染人数为因变量，根据 R 型聚类结果选取 4 种典型指标（这里选取患病人数、基本传染数、潜伏期、治愈率四个指标），建立因变量与自变量的数学表达式。最后根据预测结果给出相应的应对方案。

7.2 病毒抽检对策

题目 C 提到：以目前部分研究为例，感染新冠病毒的人群中，无症状感染者的比例大约为 18%~31%。因此，可以推测确诊人数高的地区无症状感染者往往较多。但是，由于不同地区医疗水平、经济状况、防疫措施等方面存在差异，单根据确诊人数推导无症状感染人数存在一定的局限性。因此，我们通过问题一确定的部分评价指标以及建立的主成分评价模型对某地区的严重程度进行综合评价。评价模型满足：

$$Z=0.4417y_1+0.2388y_2+0.1632y_3+0.0918y_4 \quad (7.1)$$

根据评价结果制定病毒抽检对策。具体抽检措施及对应的代表地区如表 7.1 所示。

表 7.1 无症状感染地区程度划分及应对措施

类别	严重程度	抽检措施	代表城市
I 类	严重	对无症状感染者开展大范围的病毒排查力度，并严格执行戴口罩等有关措施；	湖北
II 类	一般	对无症状感染者开展一定范围的病毒排查力度，并严格执行戴口罩等有关措施；	四川
III 类	轻微	由于严重程度较低，可开展小范围的检测措施，继续执行戴口罩等有关措施	西藏

7.3 预测模型建立与求解

这里主要考虑两类预测方法，一类是响应面分析法，另一类是 SEIR 模型。

7.3.1 响应面预测模型建立

我们以无症状感染人数 Y 为因变量，根据 R 型聚类结果选取 4 种典型指标：患病人数 P、基本传染数 R_0、潜伏期 T、治愈率 R_c 四个指标。为预测无症状感染人数发展趋势，我们拟采用响应面预测模型，建立 Y 与四个指标的关系式。

响应面设计方法（Response Surface Methodology, RSM）通过设计合理的试验方法并通过合理的操作得到试验结果数据，采用多元二次线性回归方法建立自变量和响应值之间的函数关系，通过回归方程来分析响应值的变化趋势，该方法在解决多变量问题中较为常见[10]。

本次建模选用全球顶尖级的实验设计软件 Design-Expert，选用 Box-Behnken Design(BBD)实验设计方法。在设计过程中，响应面指响应变量（因变量）Y 与一组输入变量 $x_1,x_2,...,x_n$ 之间的函数关系式，满足 $y=f(x_1,x_2,...,x_n)$。

在本问题中，响应变量指无症状感染人数 Y，自变量考虑了 P、R_0、T 和 R_c 四个因素。通过该软件，设计了 29 组实验，定量分析无症状感染人数与各影响因素之间的数学表达式。

由于篇幅限制，表 7.2 仅展示部分实验，完整实验涉及见附件。

表 7.2 响应面分析实验设计（部分）

项目	P	R_0	T	R_c	Y
实验一	2.75	5	12.5	0.2	2202
实验二	5	3	12.5	1	1971

项目	P	R_0	T	R_c	Y
实验三	0.5	3	20	0.6	1325
实验四	0.5	3	5	0.6	883
实验五	2.75	3	12.5	0.6	1243

7.3.2 响应面预测模型求解与分析

利用 Box-Behnken Design（BBD）方法对获取的数据进行分析，由此以无症状感染人数 Y 为响应值，以 P、R_0、T 和 R_c 四个因素为自变量，分别记为 A、B、C、D，建立了响应面二次多项式，具体如式 7.2 所示。

$$y = 1.65+0.35A+0.57B+0.10C-0.55D-0.16AB-0.11AC+0.15AD \\ -0.04BC+0.14BD+0.02A^2+0.09B^2-0.18C^2-0.05D^2 \quad (7.2)$$

式中，患病人数 P 单位为万人，基本传染数 R_0、治愈率 R_c 单位为 1，潜伏期 T 单位为天，无症状感染人数 Y 单位为千人。

在响应面分析过程中，用 F 值进行统计结果的显著性检测，利用 p 值来检测回归系数的显著性，p 值越小，表明结果越显著。表 7.3 展示了部分响应面分析结果，具体分析结果见附件。由表可得，模型 F 值为 5.37，$p<0.05$，说明模型具有显著的适应性，回归方程中各因素与响应值之间的关系是显著的，即该模型可信度较高，可对无症状感染人数进行较好的预测。

表 7.3 响应面分析表（部分）

参数	平方和	自由度	均方	F 值	p 值
模型	9.84	14	0.70	5.37	0.0017（显著）
A	1.48	1	1.48	11.29	0.0047
B	3.91	1	3.91	29.93	< 0.0001
C	0.13	1	0.13	0.98	0.3387
D	3.68	1	3.68	28.10	0.0001
残差	1.83	14	0.13		
净误差	0.23	4	0.059		-
总误差	11.67	28			

一般情况下，在对数据进行分析时，残差正态分布概率越接近直线，残差与方程预测值对应关系越混乱，模型可信度越高。通过分析，本模型得到相关对应关系如图 7.1-图 7.4 所示。

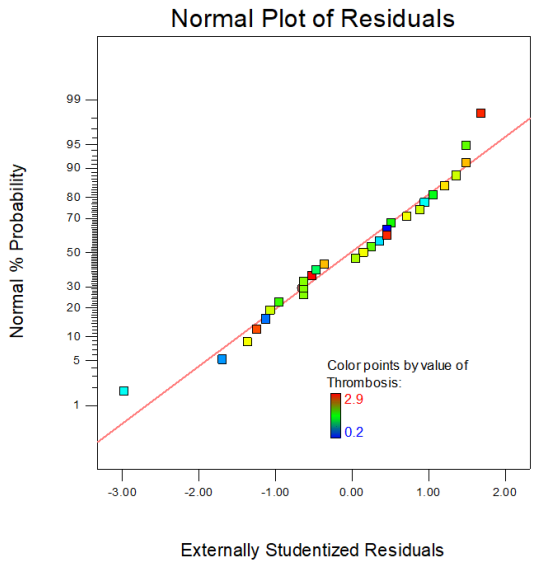


图 7.1 残差正态概率分布图

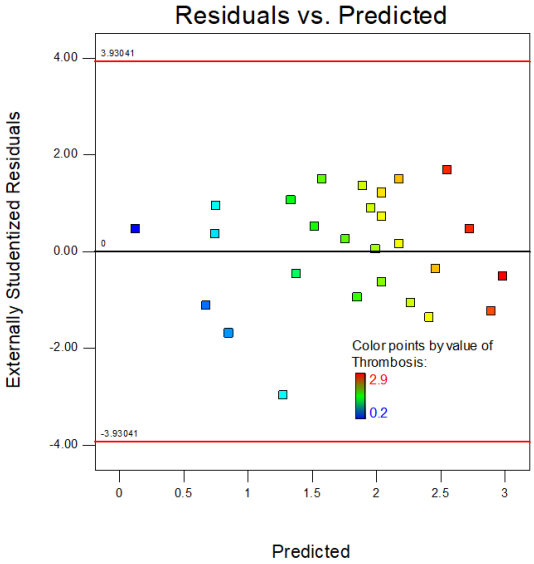


图 7.2 Residual vs predict 分布图

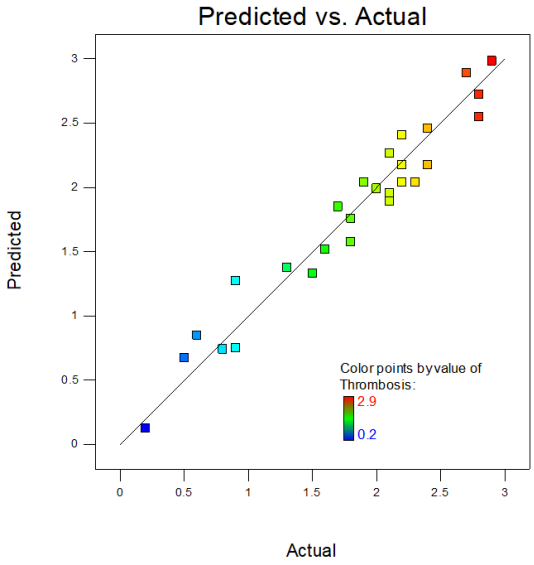


图 7.3 Predict vs Actual 分布图

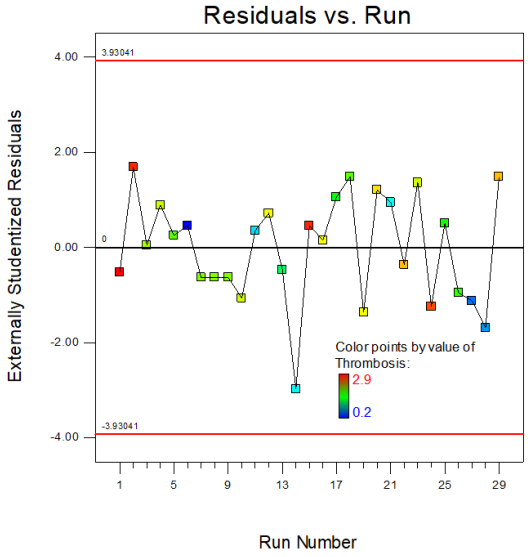


图 7.4 Residuals vs Run 分布图

由图 7.1-图 7.4 可以看出，本模型得到的残差正态概率分布曲线基本接近直线段，分布合理，残差与方程预测对应关系图分散度高，通过响应面二次多项式计算得到的预测值与真实值基本靠近同一条直线，说明利用此方法得到的模型具备一定的准确度。

接下来绘制双因素关系图对各个因素展开分析，如图 7.5-图 7.10 所示。

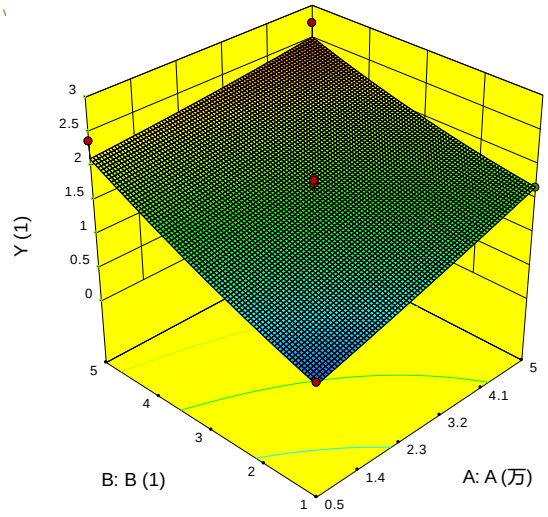


图 7.5 因素 AB 关系图

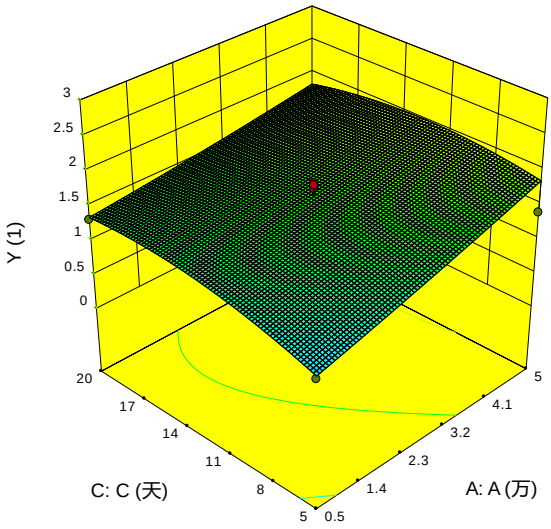


图 7.6 因素 AC 关系图

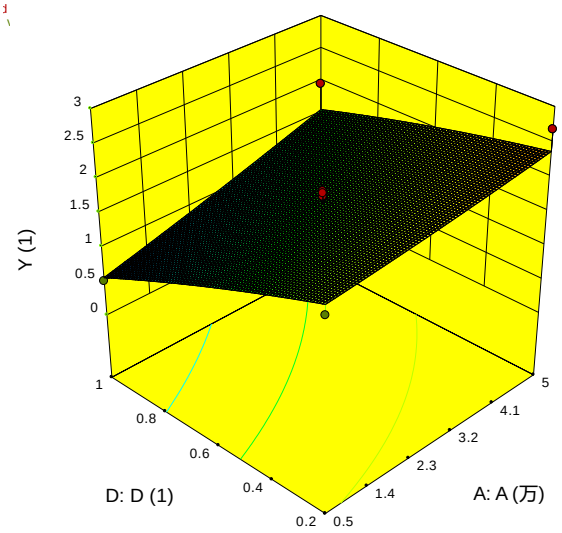


图 7.7 因素 AD 关系图

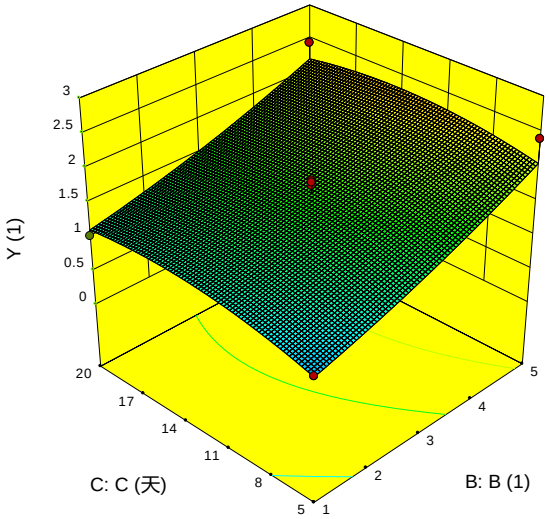


图 7.8 因素 BC 关系图

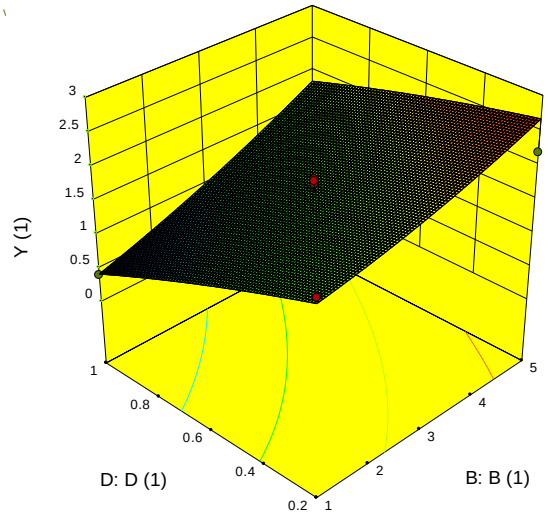


图 7.9 因素 BD 关系图

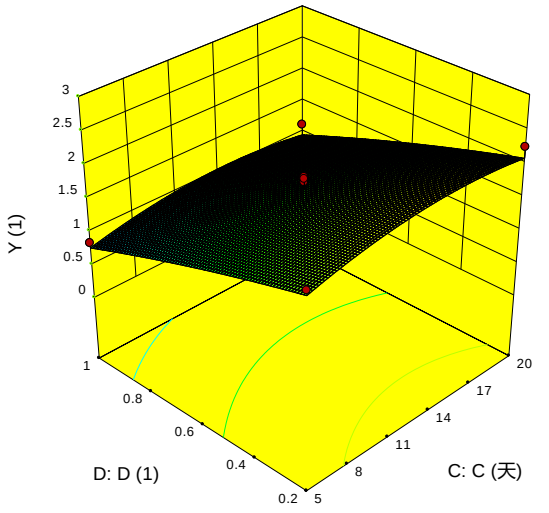


图 7.10 因素 CD 关系图

由图 7.5-图 7.10 可知，无症状感染者人数 Y 与患病人数 P、基本传染数 R_0、潜伏

期 T 呈正相关，与治愈率 R_c 呈负相关。这符合我们的认识，即患病人数越多，由于基数较大，也会对应着较多的无症状感染者；基本传染数越多，病毒传播能力越强，危害越大；潜伏期越长，病毒所带来的潜在威胁越大，也会导致较多的无症状感染者；而对于治愈率，该参数越高意味着越多的人体内携带抗体，患病的人数越少，因而对应的无症状感染者也越少。通过式 7.2 以及前面的分析，可得到四种因素对无症状感染者的敏感性（影响大小）满足：基本传染数 R_0＞治愈率 R_c＞患病人数 P＞潜伏期 T。这一结论可在我们制定相应的措施时提供数学依据。

式（7.2）给出了无症状感染者的预测公式，该数学表达式中能够改变的影响因素有治愈率（R_c）和基本传染数（R_0）。前者主要通过借助医学手段提高病人的治愈率，后者主要是通过一定的手段降低传播性（如强制隔离、佩戴口罩等）。

我们以湖北省为例，分析无症状感染者的发展规律。由于目前官方给出的无症状数据量较小，不适合通过时间序列模型对未来发展趋势展开预测。因此，我们主要通过调整治愈率（R_c）和基本传染数（R_0），分析未来无症状感染者的发展趋势。图 7.11 展示了湖北省近 20 天无症状患者变化趋势。可以看出，若继续执行当前的防控措施，后期无症状感染趋势会逐渐减少，直至新增数为 0。

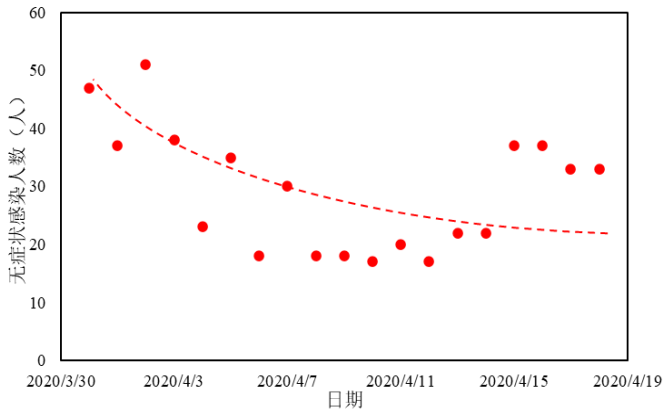


图 7.11 湖北省近 20 天无症状感染者变化趋势

对于后期预测，我们考虑四种方案：（1）继续执行当前防控措施，即各参数保持不变；（2）完全取消防疫限制（如隔离、佩戴口罩），即基本传染数提高一倍；（3）继续封城，严格限制外出，即基本传染数降至 1；（4）提高治愈率 15%。分析四种情况下未来无症状感染人数发展趋势。预测结果如图 7.12 所示。

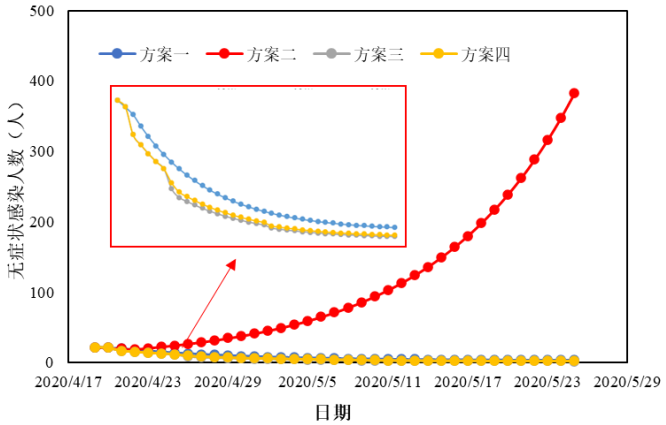


图 7.12 四种情况下湖北省未来无症状感染人数预测

可以看出，如果不加以控制，未来无症状感染者人数会陡增；若继续执行当前管控

措施，无症状感染者会持续下降，预计在五月底每天新增的无症状感染者会降至 0；若按照方案三、方案四执行，后期无症状感染者也会迅速下降，且下降幅度大于方案一。然而，由于方案三需要进一步牺牲经济，方案四需要大量医疗投入，这两种方案实施难度大、成本高，难以推广。因此，对于湖北地区，应继续保持当前防控措施，加大口罩佩戴宣传力度，对于无症状感染者继续保持医学观察，对于转正确诊案例及时隔离，避免二次传播。

7.3.3 修正 SEIR 预测模型建立

响应面预测模型是从统计学方面对问题展开分析的，这里我们从病毒传播模型角度展开分析。由于 COVID-19 病毒还具有潜伏期，因此，这里选用考虑潜伏期的传染病模型 SEIR。该模型基本流程图满足图 7.13。

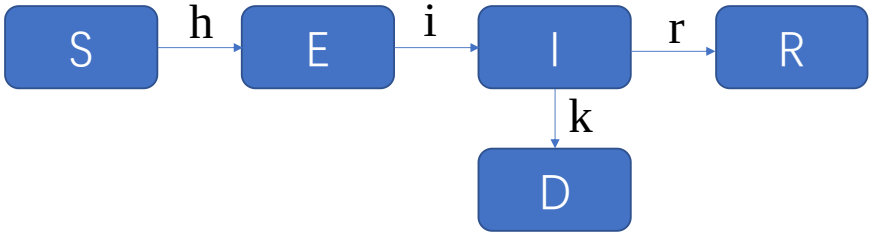


图 7.13 SEIR 基本流程图

其中，S 为易感人群；I 为患病者；D 为死亡患者； E 为潜伏者；R 为康复者；h 为感染力，表示每位患者平均每日可感染人数，h=$\alpha\beta$，其中α为接触率，即平均每个人接触到患者的概率，β为感染效率，即被感染的概率；i 为转病率；r 为治愈率；k 为病死率。

建模微分方程组[11]：

易感者可表示为：

$$dS/dt=-h(E+S)\Rightarrow S_{t+1}=S_t-h(E_t+I_t) \tag{7.3}$$

潜伏者可表示为：

$$dE/dt=h(E+I)-\mathrm{iE}\Rightarrow E_{t+1}=E_t+h(E_t+I_t)-iE_t \tag{7.4}$$

患病者可表示为：

$$dI/dt=iE-rI-kI\Rightarrow I_{t+1}=I_t-rI_t-kI_t+iE_t \tag{7.5}$$

康复者可表示为：

$$dR/dt=rI\Rightarrow R_{t+1}=R_t+rI_t \tag{7.6}$$

病死患者可表示为：

$$dD/dt=kI\Rightarrow D_{t+1}=D_t+kI_t \tag{7.7}$$

这种情况适用于以下条件：现阶段不加以任何应对措施，即自限度治疗，并忽略人口出生、死亡等影响。显然，这种情况比较理想，一是以为我们开展了相应的应对措施（戴口罩、封城等），减少了易感人群数 S；二是潜伏者转阴（这里取 14 天），导致患病者相对减少。因此对该模型进行了修正，修正结果如下：

易感者可表示为：

$$S_{t+1}=S_t-r\beta I_t-r_2\beta_2E_t+\beta_3E_{t\text{-}14} \tag{7.8}$$

潜伏者可表示为：

$$E_{t+1} = E_t + r\beta I_t + r_2\beta_2 E_t - \beta_3 E_{t-14} \tag{7.9}$$

患病者可表示为：

$$I_{t+1} = I_t + \alpha E_t - (\gamma + k) I_t \tag{7.10}$$

康复者可表示为：

$$R_{t+1} = R_t + yI_t \tag{7.11}$$

病死患者可表示为：

$$D_{t+1} = D_t + kI_t \tag{7.12}$$

式中，α 为潜伏者转阳率；β 为传染概率（接触者、患者）；r 为平均每个病人每天接触的人数；γ 为患者康复概率；β_2 为传染概率（接触潜伏者）；β_3 为潜伏者的转阴率；r_2 为潜伏者平均每天接触的人数；k 为患者死亡率。

7.3.3 修正 SEIR 预测模型求解与分析

同样以湖北省为例，其 SEIR 预测结果如图 7.14 所示。

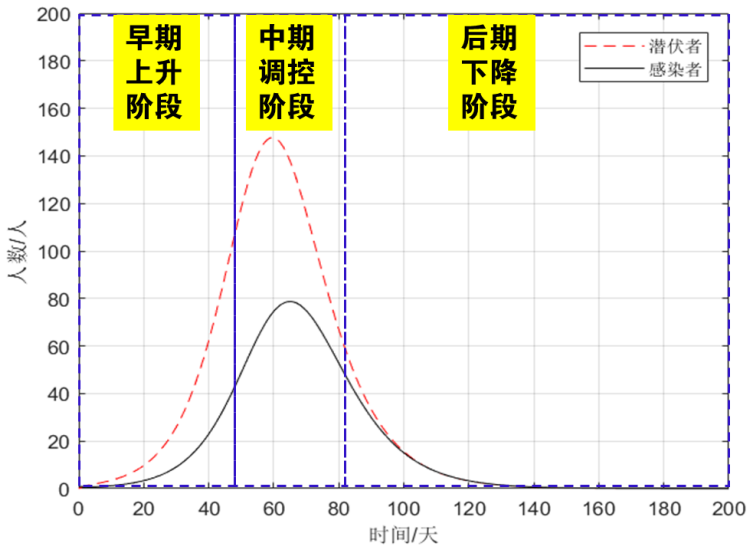


图 7.14 SEIR 模型下湖北地区无症状感染者变化趋势

可以看出，该曲线共分为三个阶段。第一阶段是早期上升阶段：属于病毒暴发早期，潜伏者（无症状感染人数）人数快速上升，由无症状感染者转阳人数也持续上升，若不加以控制，后期上升趋势还将持续进行（如图 7.15）；第二阶段为中期调控阶段，该阶段一开始实行了封城隔离、佩戴口罩、治疗等措施，无症状人数虽有上升，但是后期得到了控制，大约 22 天后达到了峰值；第三阶段为后期下降阶段，该阶段由于前面的措施持续进行，无症状感染人数下降趋势明显，在疫情开始后的 140 天内，日新增无症状感染者人数降为 0，也就是说，大约到 5 月中旬，湖北地区的新增无症状感染人数会趋近于 0，这与前面的响应面预测结果一致。

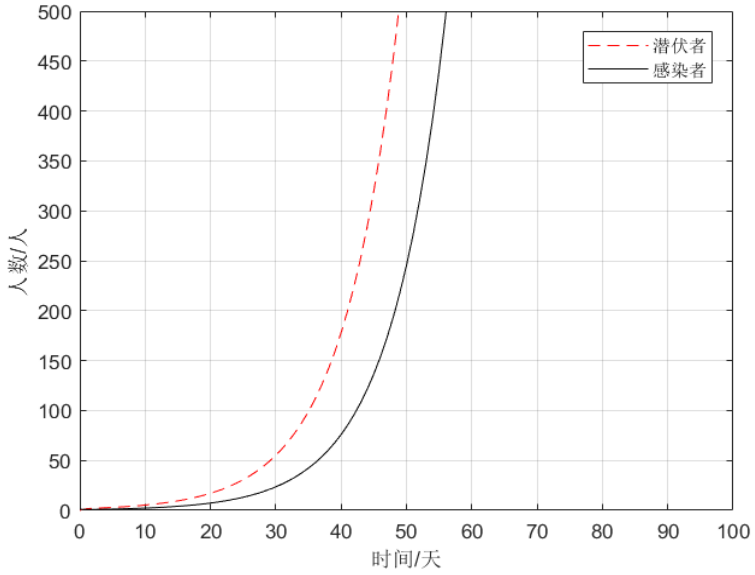


图 7.15 自然发展下无症状人数（潜伏者）预测

比较图 7.14、图 7.15 我们可以看出，湖北地区的疫情应对措施合理，极大的抑制了疫情的传播。因此，对于后期防控，我们认为应继续保持当前的防控措施，鼓励市民减少外出，多戴口罩，避免交叉感染。在后期，由于湖北地区发展态势较好，可逐渐复工复产，在稳定疫情的前提下推动经济发展。

7.4 问题二小结

本问题主要首先根据问题一中的聚类分析结果以及建立的主成分评价模型对不同地区的感染程度进行分级，然后根据分级结果给出相应的抽检对策。

在对无症状感染者进行预测时，分别从统计学和流行病学两个方面展开分析。统计学方法，我们选用了响应面预测模型，依据模型预测结果，可得到四种因素对无症状感染者的敏感性（影响大小）满足：基本传染数 R_0＞治愈率 R_c＞患病人数 P＞潜伏期 T，同时后期无症状感染者会持续下降，预计在五月下旬会趋近于 0；流行病学方面，选用了修正 SEIR 模型，该模型能够考虑潜伏期带来的影响。通过修正 SEIR 预测发现，湖北地区的防疫措施合理，若不加任何措施，后期患者数量会急剧增加，但是执行一定防疫措施后，增长趋势得到了控制。预测结果表明，疫情暴发 140 后，日新增无症状感染者人数降为 0，也就是说，大约到 5 月中旬，湖北地区的新增无症状感染人数会趋近于 0，这与前面的响应面预测结果基本一致，说明了预测模型合理，预测结果准确性较高。

7.5 模型敏感性分析

在响应面分析中，我们选择了二次插值，该方法具有较高的精度。在本节中，我们使用不同的插值方法来分析模型对插值方法的敏感性。每种方法的预测值与真实值的拟合结果如图 7.16-图 7.19 所示。

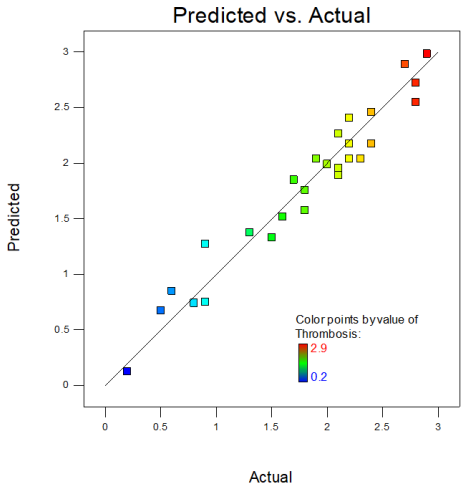


图 7.16 二次插值

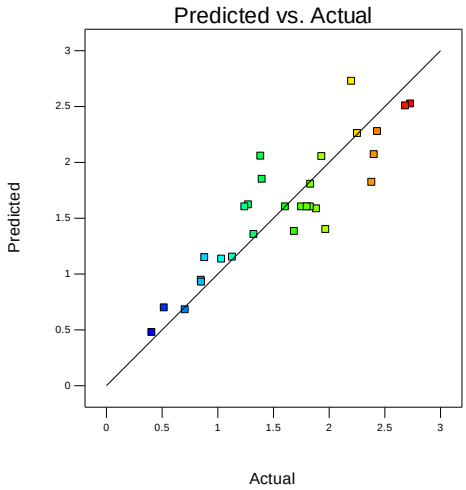


图 7.16 线性插值

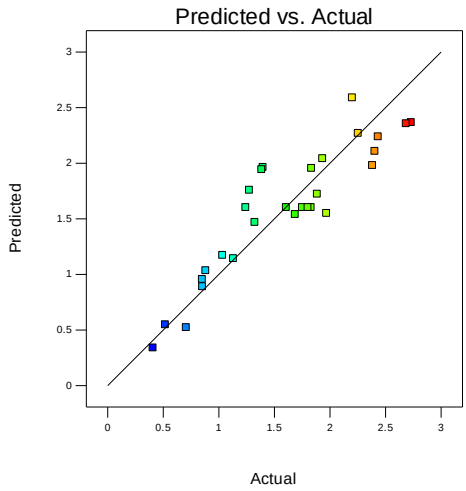


图 7.18 2FI 插值

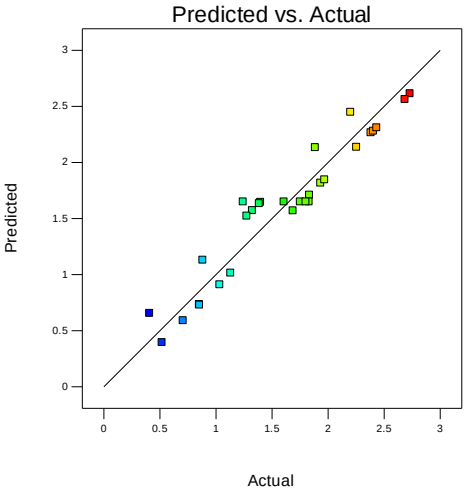


图 7.19 三次插值

通过比较可以发现，不同插值方法的拟合存在差异。随着插值多项式的增加，预测精度逐渐提高，但当插值多项式大于 3 倍时，模型精度对插值方法不敏感。如图 7.19 所示，当插值多项式为三次时，拟合效果很好，但这种方法的缺点是多项式项太多，计算繁琐。对于这个问题，二次插值可以满足要求，模型适应性较强。

八、给世卫组织(WHO)的一封信

尊敬的世卫组织工作者们：

你们好！

2020 年年初，新冠肺炎 2019-nCoV 的暴发给全世界人民带来了沉重的灾难。面对来势汹汹的肺炎疫情，你们挺身而出，参与到这场没有硝烟的战争中，积极深入疫情调研、搜集数据，并将你们的结论通告了全世界，为全人类的生存和发展敲响了紧钟。向你们致敬。

经过我们团队多天的奋斗，我们对该病毒的危害、传播特点以及对策有了较清晰的认识。我们也非常愿意将我们的认识与结论同你们分享，以便早日打赢这场疫情反击战。

首先，基于大数据背景，我们统计了 16 类传播程度不同的流行疾病，并选取 14 个指标对流行病的传播特点进行评价。我们借助数学方法，利用计算机手段，对这 16 类

流行病划分成了散发（sporadic），暴发（outbreak）、流行（epidemic）和大流行（pandemic）。四类，并用数学方法划分了不同类型之间的界限。根据评价结果，我们一致认为新冠肺炎 2019-nCoV 属于“大流行”级别传染病，这与你们的判断一致，也验证了你们决策的合理性。

然后，我们利用数学评价模型对不同地区的疫情验证程度进行了分级，并给出了相应的应对措施。对于疫情严重地区，如中国湖北省，应该对无症状感染者开展大范围的病毒排查力度，并严格执行戴口罩等有关措施；对于疫情一般严重地区，如中国四川省，应该对无症状感染者开展一定范围的病毒排查力度，并严格执行戴口罩等有关措施；对于疫情不严重地区，如中国西藏省，可开展小范围的检测措施，并继续执行戴口罩等有关措施。在疫情未来趋势预测方面，我们以中国湖北省为例，从统计学和传染病学两个方面对未来疫情发展趋势就行了预测分析。结果显示，对于疫情暴发地，如果不采取任何措施，短时间内，感染人数会呈现指数级上升，到后期难以控制；中国湖北自 1 月 24 日陆续采取严格控制措施后，如封城、限制居民外出、要求进入公众场合佩戴口罩等，在大约 22 天后疫情达到了顶峰，之后患病人数/无症状感染人数呈下降趋势。初步预计，按此措施执行，到五月中下旬湖北地区新增无症状感染人数将会将为 0，疫情基本得以控制。这为世界其它国家和地区制定应对措施提供了参考。

最后，我由衷的希望你们继续宣传疫情防控措施：包括（1）隔离确诊者；（2）跟踪观察无症状感染者；（3）呼吁政府限制居民外出，并要求进入公共场合佩戴口罩；（4）必要时可以封城。

希望在我们共同的努力下，全世界人民能够早日摆脱 2019-nCoV 的影响，步入崭新的生活篇章。

此致。

2020/4/19

九、模型评价与推广

9.1 模型评价

（1）在对“流行” (Epidemic) 和“大流行”(Pandemic) 病进行界限划分时，我们参考了流行病学的基本分类结果，并利用 R 型聚类分析法分析了 14 项影响流行病传播的因素，考虑全面；同时，我们挑选了 16 类流行病，利用 Q 型聚类分析法将其聚成散发（sporadic），暴发（outbreak）、流行（epidemic）和大流行（pandemic）4 大类，该分类结果与 WHO 历年的评价结果一致。

（2）我们选用了主成分分析法对四大类中的流行病进行综合打分，根据打分结果对四大类传播级别进行定量划分，该方法减少了主观因素的影响，分类结果比较合理。

（3）在进行无症状感染者预测时，我们分别从统计学和传染病学两个方面展开分析。前者选用了响应面分析法，后者选择了考虑潜伏期存在的 SEIR 模型，两种方法成功实现未来无症状感染者的预测。

9.2 模型改进

（1）在进行聚类分析时，由于时间有限，仅选择了 16 类传染病的相关数据，后续优化时有待完善。

（2）在进行无症状感染者未来发展趋势预测时，由于数据获取难度较大，我们仅对湖北地区展开了分析，后续可开展多地区分析。

参考文献

[1]梁桂珍,郝林莉.一类潜伏期和染病期均传染的 SEIQR 流行病模型的稳定性[J].西南师范大学学报(自然科学版),2020,45(03):1-9.

[2]韦宵宵. 带有流动人口肺结核模型的研究[D].北京建筑大学,2017.

[3]王绍凯. 几个传染病模型复杂动力行为研究[D].哈尔滨工业大学,2010.

[4]王茜. 两类传染病模型动力学分析[D].中北大学,2018.

[5]郭晓霞. 两类具有时滞的 HIV 模型的研究[D].山西师范大学,2017.

[6]石耀霖,程惠红,黄禄渊,任天翔.用离散随机模型研究湖北新冠肺炎 COVID-19 流行病动力学特征[J].中国科学院大学学报,2020,37(02):145-154.

[7]王娜. 麻风风险预测模型的构建与效能评价[D].山东大学,2018.

[8] 赵海龙 ,张丹丹 ,黄松 ,莫石 ,魏浩 .基于皮尔逊相关系数的海南省地闪密度与雷击故障关系分析.

[9]司守奎,孙兆亮.数学建模算法与应用,北京:国防工业出版社,384-385,2015.

[10]卜庆状,郝晓莉,张馨予,陈芳,李丽娜.响应面试验优化矿物源腐植酸肥料中可溶性腐植酸的快速测定方法[J].辽宁农业科学,2019(06):15-18.

[11]王园园. 具有时滞的 SEIR 的肺结核模型研究[D].西安科技大学,2012.

附 录

1. 响应面预测模型

（1）实验设计

实验	A	B	C	D
1	2.75	5	12.5	0.2
2	5	3	12.5	1
3	0.5	3	20	0.6
4	0.5	3	5	0.6
5	2.75	3	12.5	0.6
6	2.75	3	20	1
7	2.75	1	12.5	1
8	5	3	12.5	0.2
9	2.75	1	20	0.6
10	0.5	3	12.5	0.2
11	5	3	20	0.6
12	2.75	5	5	0.6
13	2.75	3	12.5	0.6
14	2.75	3	12.5	0.6
15	2.75	3	20	0.2
16	5	3	5	0.6
17	2.75	5	20	0.6
18	5	1	12.5	0.6
19	2.75	1	12.5	0.2
20	0.5	3	12.5	1
21	2.75	3	12.5	0.6
22	2.75	1	5	0.6
23	2.75	3	12.5	0.6
24	5	5	12.5	0.6
25	2.75	5	12.5	1
26	0.5	5	12.5	0.6
27	2.75	3	5	1
28	0.5	1	12.5	0.6
29	2.75	3	5	0.2

（2）预测结果

	Sum of		Mean	F	p-value		
Source	Squares	df	Square	Value	Prob > F		
Model	9.839911	14	0.702851	5.374628	0.001665	significant	
A-A	1.476307	1	1.476307	11.28917	0.004671		
B-B	3.913634	1	3.913634	29.92716	8.25E-05		
C-C	0.12834	1	0.12834	0.981403	0.338667		
D-D	3.67524	1	3.67524	28.10419	0.000112		
AB	0.099856	1	0.099856	0.763589	0.39695		
AC	0.051529	1	0.051529	0.394037	0.540289		

AD	0.0897	1	0.0897	0.685929	0.421448		
BC	0.0057	1	0.0057	0.043589	0.837627		
BD	0.076176	1	0.076176	0.58251	0.457999		
CD	0.0004	1	0.0004	0.003059	0.956676		
A^2	0.003886	1	0.003886	0.029713	0.865611		
B^2	0.048505	1	0.048505	0.370916	0.55226		
C^2	0.206153	1	0.206153	1.576434	0.229827		
D^2	0.013296	1	0.013296	0.101674	0.754537		
Residual	1.830808	14	0.130772				
Lack of Fit	1.596277	10	0.159628	2.722503	0.173378	significant	
Pure Error	0.234531	4	0.058633				
Cor Total	11.67072	28					

2.聚类分析代码

（1）R 型聚类分析代码

```
clc,clear
a=load('gj.txt');   %把原始数据保存在纯文本文件 gj.txt 中
b=zscore(a); %数据标准化
r=corrcoef(b); %计算相关系数矩阵
% d=tril(1-r); d=nonzeros(d)'; %另外一种计算距离方法
d=pdist(b','correlation'); %计算相关系数导出的距高
z=linkage(d,'average'); %按类平均法聚类
h=dendrogram(z); %画聚类图
set(h,'Color','k','Linewidth',1.3) %把聚类图线的颜色改成黑色，线宽加粗
T=cluster(z,'maxclust',7); %把变量划分成 6 类
for i=1:7
tm=find(T==i); %求第 i 类的对象
tm=reshape(tm,1,length(tm)); %变成行向量
fprintf('第%d 类的有%s\n',i,int2str(tm)); %显示分类结果
axis([0,15,0,1.2])
end
```

（2）Q 型聚类分析代码

```
clc,clear;
load gj_1.txt %把原始数据保存在纯文本文件 gj.txt 中
a=zscore(gj_1);                    %数据标准化处理
y=pdist(a,'cityblock');     %求 a 的行向量之间的绝对距离
yc=squareform(y);              %变换成距离方阵
z=linkage(y);                     %产生等级聚类图
[h,t]=dendrogram(z);          %画聚类图
set(h,'Color','k','Linewidth',1.3) %把聚类图线的颜色改成黑色，线宽加粗
T=cluster(z,'maxclust',9);%把对象分成 3 份，参数可自行修改成 2,4,5 等，记得将下一行 i 值修改
```

```
for i=1:16;
    tm=find(T==i);              %求第 i 类的对象
    tm=reshape(tm,1,length(tm));%变成行向量
    fprintf('第%d 类的有%s\n',i,int2str(tm));%显示分类结果
    axis([0,16,1,9])
end
```

3.主成分分析代码

```
clc,clear
gj=load('input.txt');     %把原始数据保存在纯文本文件 input.txt 中
gj=zscore(gj); %数据标准化
r=corrcoef(gj);   %计算相关系数矩阵
%下面利用相关系数矩阵进行主成分分析，x 的列为 r 的特征向量，即主成分的系数
[x,y,z]=pcacov(r) %y 为 r 的特征值，z 为各个主成分的贡献率
Z = cumsum(z);
plot(Z,'*-r')
f=repmat(sign(sum(x)),size(x,1),1); %构造与 x 同维数的元素为±1 的矩阵
x=x.*f %修改特征向量的正负号，每个特征向量乘以所有分量和的符号函数值
num=4;   %num 为选取的主成分的个数
df=gj*x(:,[1:num]);   %计算各个主成分的得分
tf=df*z(1:num)/100; %计算综合得分
[stf,ind]=sort(tf,'descend');   %把得分按照从高到低的次序排列
stf=stf'; ind=ind';
```

4. SEIR 模型代码

```
%SEIR 模型
clear;clc;
%参数设置
N=60000000;
I=1;%传染者
R=0;%康复者
D=0;%死亡患者数量
E=0;%潜伏者
S=N-I;%易感染者
r=2;%接触病患的人数
a=0.125;%潜伏者患病概率
B=0.8;%感染概率
y=0.143;%康复概率
B2=0.03;%接触潜伏者
B3=0.859;%转阴率
r2=10;%潜伏者每天接触的人数
k=0.025373;%死亡率
T=0:200;
```

```
% for idx =1:length(T)-1
%         S(idx+1)=S(idx)-r*B*I(idx)*S(idx)/N;%易感人数迭代
%         E(idx+1)=E(idx)+r*B*S(idx)*I(idx)/N-a*E(idx)%潜伏者人数迭代
%         I(idx+1)=I(idx)+a*E(idx)-(y+k)*I(idx);%患病人数迭代
%         R(idx+1)=R(idx)+y*I(idx);%康复人数迭代
%         D(idx+1)=D(idx)+k*I(idx);%死亡患者人数迭代
% end
%修正
for idx =1:length(T)-1
    S(idx+1)=S(idx)-r*B*I(idx)-r2*B2*E(idx)+B3*E(idx);%易感人数迭代
    E(idx+1)=E(idx)+r*B*I(idx)+r2*B2*E(idx)-B3*E(idx);%潜伏者人数迭代
    I(idx+1)=I(idx)+a*E(idx)-(y+k)*I(idx);%患病人数迭代
    R(idx+1)=R(idx)+y*I(idx);%康复人数迭代
    D(idx+1)=D(idx)+k*I(idx);%死亡患者人数迭代
end
%plot(T,S,T,E,T,I,T,R,T,D);
plot(T,E,'--r',T,I,'-k');
grid on;
xlabel('日期');
ylabel('人数/人');
%legend('易感者','潜伏者','潜伏者转阳','康复者','死亡者');
title('SEIR 模型');
%plot(T,E,T,I,T,R,T,D);
plot(T,E,'--r',T,I,'-k');
grid on;
xlabel('时间/天');
ylabel('人数/人');
axis([0 100 0 500]);
%legend('潜伏者','感染者','康复者','死亡者');
legend('潜伏者','感染者');

set(gca,'ytick',[0:50:500]);
%title('疫情情况');
```