第二届“数学中国杯”数学建模网络挑战赛

承 诺 书

我们仔细阅读了第二届“数学中国杯”数学建模网络挑战赛的竞赛规则。

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的, 如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们将受到严肃处理。

我们允许数学中国网站(www.madio.net)公布论文，以供网友之间学习交流，数学中国网站以非商业目的的论文交流不需要提前取得我们的同意。

我们的参赛报名号为：

参赛队员 (签名) ：

队员 1：

队员 2：

队员 3：

参赛队教练员 (签名)：

参赛队伍组别：

第二届“数学中国杯”数学建模网络挑战赛

编 号 专 用 页

参赛队伍的参赛号码：（请各个参赛队提前填写好）：

#1036

竞赛统一编号（由竞赛组委会送至评委团前编号）：

竞赛评阅编号（由竞赛评委团评阅前进行编号）：

第二届数学中国数学建模网络挑战赛 #1036

2009年 第二届“数学中国杯”
数学建模网络挑战赛

题　　目 基于流感疫苗改装研究

关 键 词 自回归移动平均模型、扩散权重、聚类分析、优先级系数

摘　　要

本文就病毒亚型分布与爆发情况，以及流感监测网络分布概况进行了分析，探讨了疫苗改装、下一年优势毒株的可能分布区域与需援助国家的优先级问题。

问题一，由地理区域相似性将全球划分成七个区域，根据每个区域$H1$、$H3$、B、A四种病毒亚型爆发情况与人口比重进行权值计算求得每种病毒亚型的扩散权重，然后引入ARIMA(自回归移动平均模型)对数据进行统计分析，得到每种病毒亚型在每个地区的独立扰动值，进而预测出每种病毒亚型在各区域流行季节中所占比重，依此可得改装疫苗的投放方案（表 4-2）。对于疫苗改装的可行性，我们引进覆盖率的标准进行了评价。（表）

问题二，通过对七个区域病毒亚型扩散权重的计算与分析，我们发现对于流行毒株种类的变更主要出现在冬季的后期，而对于下一年威胁较大的病毒可能出现在北美洲和亚洲地区。

问题三，通过引入聚类分析中的分层聚类，对193个国家根据其卫生支出占GDP百分比聚类，将其分成8个类别（表4-4）。通过计算国家优先级系数γ，确定需援助国家优先级。系数γ越大，国家优先级越高，越需要联合国经费援助。

参赛队号 #1036

参赛密码 _____
（由组委会填写）

所选题目 B题

第二届数学中国数学建模网络挑战赛 #1036

Abstract

一、问题重述

流感是一种广泛流行于世界范围内的疾病，每次流感大流行都会造成多人死亡和巨大损失。世界卫生组织大力推荐将疫苗作为一种有效的预防措施来抗击这种潜在的致命性疾病。如果疫苗毒株1和流行的病毒类型相匹配，那么大约有50%-80%的疫苗接种者能够抵抗流感的侵袭。即使疫苗不能完全抵御流感的侵袭，它也可以降低流感发病的严重程度以及严重并发症的发生率。但流感疫苗所能产生的抗体是短效的，所以每年流感流行季节到来前，都需要重新接种疫苗。每年冬天是流感的流行季节，在流行季节到来前1-2个月接种疫苗，能达到较为良好的防护效果。

流感病毒分为A、B、C型。A型病毒容易发生变异，根据抗原不同可区分不用亚种，B型病毒变异缓慢，C型病毒甚少对人类造成威胁。对于病毒变异的多样性和变异性，每年采用疫苗的成分不同世界卫生组织专家痛过对全球疫情的监控来收集数据，在每年二月份预测新的流行季节中流感流行情况，并确定毒株品种作为新年度北半球流感疫苗的推荐成分。需要给药品制造商留出半年左右的时间以生产和投放市场。现在管用的推荐方案是三联装一面，也就是每份疫苗中有三种经过灭活或裂解处理的毒株，分别问两个A型和一个B型。

根据流感传播、构成、变异与疫苗市场建立模型解决以下问题：

【问题1】：降低流感疫苗的制造成本，将三联装的流感疫苗改装成双联装。将北半球和南半球分别划分成稍小的区域，并使用不同的疫苗针对不同的区域来投放，请建立的模型，设计一个可行的投放方案，并设计一个评估标准来评估其效果，使之能与现行方案进行对比评价。

【问题2】：考虑选择具体毒株，评估和预测对下一年威胁性最大的病毒品种。建立合理的模型，在流感流行记录中，筛选出对下一年威胁较大的病毒出现的区域。

【问题3】：流感病毒依靠多种途径传播。一旦出现流行，就难以预测和控制其最终的流行范围和持续时间。所以为了控制流感的传播，需要有快速反应的应对措施。世界卫生组织（WHO）在世界范围内建立了流感监测网络。包括世界流感中心及实验室，国家级流感中心，数量众多的国家级流感监测网络实验室和哨点医院。流感监测网络的作用是监视流感在全球的活动情况，并且及时准确地将流感爆发的信息、病毒分离的结果上报世界卫生组织，进行病毒的研究鉴定。为了提高流感监测网络的覆盖率和反应能力，一方面需要提高各级流感中心对病毒样本的研究鉴定能力，另一个很重要的方面是需要增加基层投入，主要包括流感监测网络实验室和哨点医院的数量。假设联合国有部分经费可以援助不同国家，用以建设流感监测网络实验室和哨点医院。请依据数据，建立合理的模型，评估需要援助的国家或地区的优先级。

二、符号说明与问题假设

2.1 符号说明：

2.1.1 问题一、二的符号说明

$H1$ 流感病毒 A 的主要亚型之一

$H3$ 流感病毒 A 的主要亚型之二

B 流感病毒 B 的全体亚型

A 流感病毒 A 除 $H1$ 与 $H3$ 外其他亚型

V_{ijk} 第 k 个区域中第 i 个国家在该月病毒 j 的扩散权值因子

n 区域病毒 j 的扩散权值因子的指数

M_{ki} 第 k 个区域中第 i 个国家的人口数

M_k 第 k 个区域中人口总数

D_{ki} 第 k 个区域中第 i 个国家的人口比例

O_{kj} 第 k 个区域病毒 j 的扩散权重

P_k 第 k 个区域人口数占世界总人口数的百分比

W_j 世界范围内第 j 种病毒扩散权重

X_{jt} 病毒 j 在权值时间序列函数中的取值

φ_t 时间序列函数的系数

ε_t 时间序列函数的残差

ϕ_t 时间序列函数中残差的系数

2.1.2 问题三的符号说明

β_i 第 i 国家哨点医院医生占为医生总数的比值为

S_{mi} 第 i 个国家中哨点医院的医生数

S_m 第 i 个国家医院的医生总数。

δ_i 哨点医院的医生密度

ϕ 标准化后哨点医院的医生密度

γ 国家的优先级系数

2.2 问题假设：

1、假设所收集的流感病毒资料真实有效。
2、假设 A 型病毒在研究区域与时段中分为 H1、H3 和其余总和三种亚型。
3、假设 B 型病毒在研究区域与时段中不发生变异，不存在亚种。
4、假设 C 型病毒在研究区域与时段中对人类不造成威胁。
5、假设全球按北美洲、南美洲、欧洲、亚洲、北非、撒哈拉以南非洲沙漠、大洋洲区域划分。

三、问题分析

3.1 问题一的分析

流感的预防与控制可通过在人体内接种疫苗类预防性生物制品方式达到目的，然而疫苗的造价高昂，需要依随病毒的变异而不断更新变化等因素的影响，使流感的预防与控制难度进一步加大。为疫苗的推广进一步普及，目前的三联装疫苗可根据病毒的地域性分布差异而改进成二联装疫苗。

A 型流感病毒可以感染多种哺乳动物（例如猪、马）和鸟类，而 *B* 型和 *C* 型主要局限于人间感染。仅 A 型和 B 型流感病毒能够致人患病。所有目前确认的 A 型流感病毒的 16 种 *HA* 和 9 种 *NA* 亚型都存在于野生的水鸟群落中。感染人类的流感病毒亚型通常是 *H*1、*H*2 或 *H*3 以及 *N*1 或 *N*2 。故在以下模型中将病毒分为 *H*1、*H*3、*B* 、*A* （*A* 型中除 *H*1、*H*3 外的其他亚型)四种类型。

对于病毒的地域性分布的相似性，我们将全球分为北美洲、南美洲、欧洲、亚洲、北非、撒哈拉以南非洲沙漠、大洋洲七大区域，对世界卫生组织流感监测项目资料中 2007 年 3 月至 2008 年 2 月数据进行如下分析：

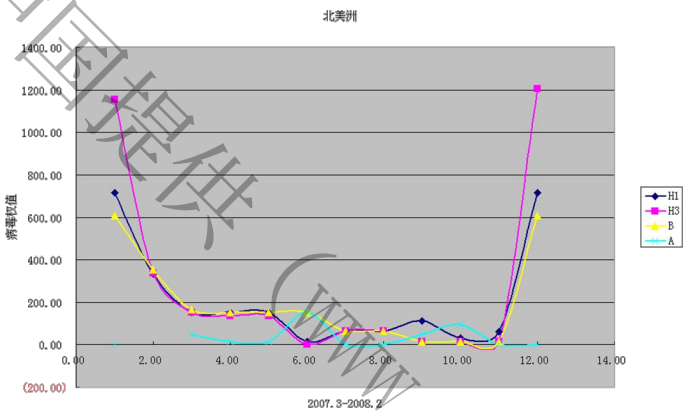


图 3-1　　北美洲 2007.3-2008.2

1）　北美洲

位于北半球，冬季为流感高发期，通过对数据的统计分析，绘制图像可以得到右图。图中可以明显看出在当年的 12 月至次年的 3 月，是流行性感冒的高发期。在高发期中 H3、H1 两种病毒有明显优势。我们通过此变化曲线预测其满足 ARIMA(Auto-regressive Integrated Moving Average Model),即自回归移动平均模型。

2）　北非

位于北半球地中海沿岸，气候，地形与欧洲有一定的相似性，但由于其地形特殊和人口密度小等原因，全年基本不存在流感高峰期，也不具有北半球流感爆发的特性，故在以下模型分析中暂时将其忽略不具体考察。

欧洲为北半球的典型模型，流感爆发时间集中，在该年的 12 月至次年的 4 月，然而其爆发期中 *H*1、*H*3、*B* 三种病毒扩散权值差值较小，在三联装改进成二联装是需要利用 *ARIMA* 模型求解。

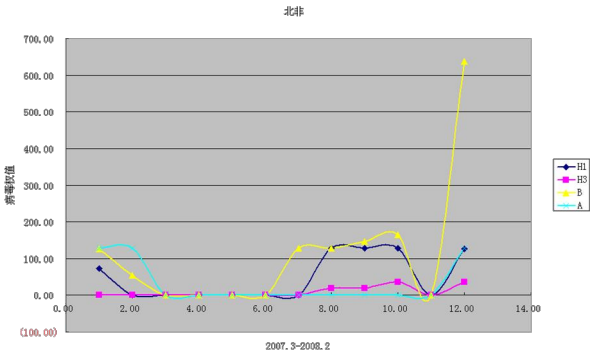


图 3-2　北非 2007.3-2008.2

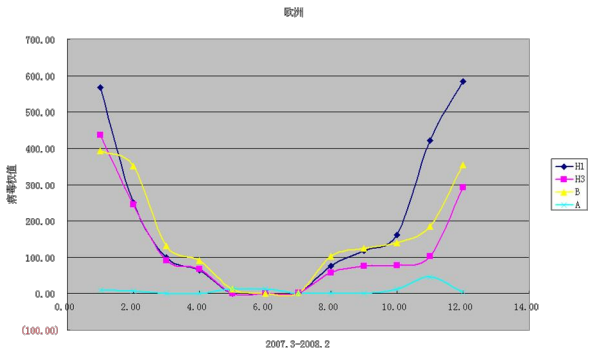


图 3-3 欧洲 2007.3-2008.2

3)　亚洲

亚洲因人数为全世界之首，其流感的影响程度大，持续时间长，而在不同时期内不同病毒的存活的差异亦相对北半球其余各州大。在模型的研究中重点的疫苗改装问题可通过病毒扩散权值因子的差异来选择适合二联装疫苗的组成。通过数值变化得到变化函数，从而推断明年的疫苗选择。

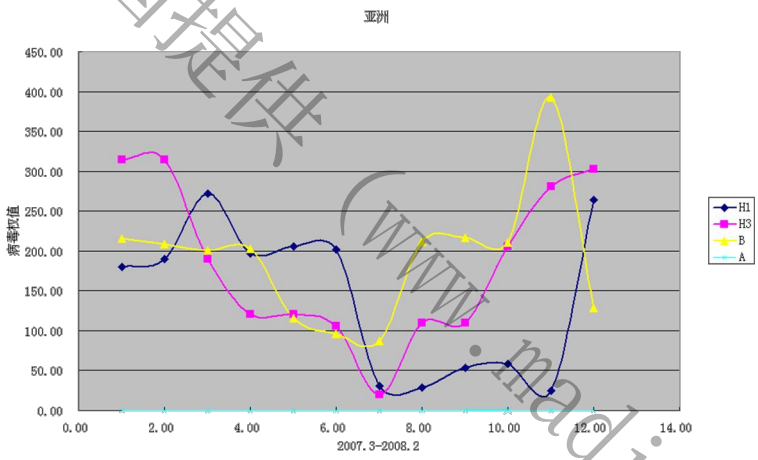


图 3-4 亚洲 2007.3-2008.2

4)　南美洲和撒哈拉以南非洲

南美洲和撒哈拉以南非洲位于东西半球相同的纬度之上流感的爆发有一定的相似性，在该年的 5 月到 8 月为高峰期，可用上述模型分析。

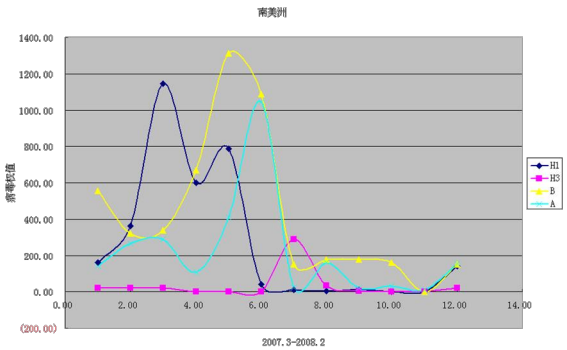


图 3-5 南美洲 2007.3-2008.2

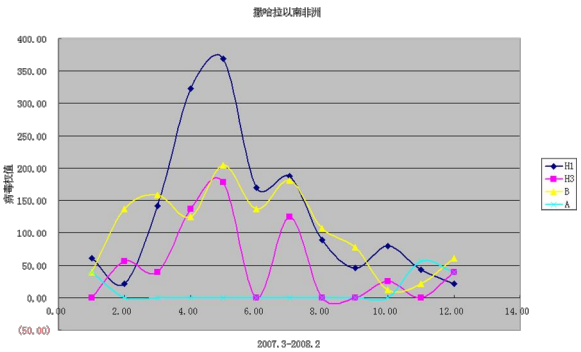


图 3-6 撒哈拉以南非洲 2007.3-2008.2

5） 大洋洲

大洋洲位于印度洋与太平洋间，人口集中且相对其他大陆较少，能收集的相关系数亦不多，在图中可分析 H3、B 两种显著地病毒。不将其作为南半球主要模型，只通过其往年数据推断明年疫苗情况。

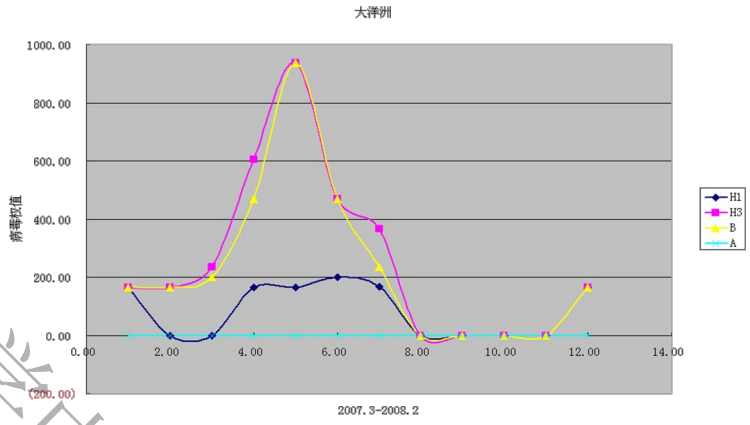


图 3-7 大洋洲 2007.3-2008.2

3.2 问题二的分析

问题二主要讨论优势毒株的选择问题。在问题一中我们从众多病毒亚型中选取最典型的四种亚型种类。由资料可得，亚型中可以根据其发现地点、时间进行编号，而成为不同的毒株。

对于评估和预测下一年的优势毒株，对2001-2008 年这 7 年的数据进行分析，并根据权值计算公式计算出各个区域内 4 中病毒亚型的权值，再将它乘以每个区域人口占世界总人口的百分比得到全球范围内的各病毒亚型所占百分比。

以 2007.3-2008.2 的实例画出的全球范围的病毒亚型百分比如下图饼状图所示，其中蓝色表示 $H1$ 亚型、褐色表示 $H3$ 亚型、绿色表示 B 亚型而紫色则表示为 A 亚型所占的比例。

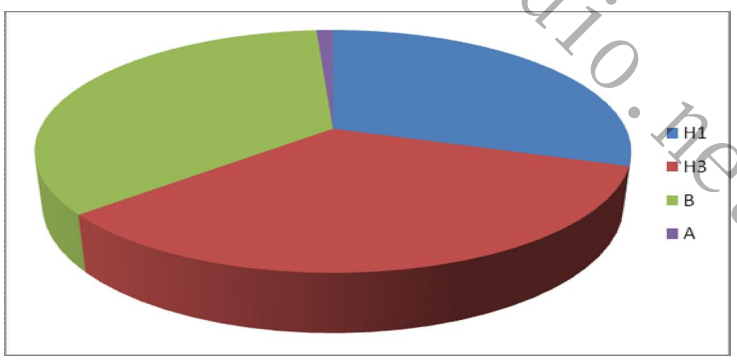


图 3-8 全球 2007.3-2008.2 病毒亚型比例

四、模型的建立与求解

4.1 问题一模型的建立与求解

为解决以下问题，于此引入 ARIMA(Autoregressive Integrated Moving Average

Model),即自回归移动平均模型。模型的定义如下：

定义 1：自回归移动平均 $ARIMA(p,d,q)$ 模型为

$$X_t = \varphi_1 X_{t-1} + \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

式中， p 和 q 是模型的自回归阶数和移动平均阶数； θ， φ 是不为零的待定系数； ε_t 是独立的误差项； X_t 是经过 d 阶差分处理后的平稳、正态、零均值的时间序列。

定义 2：在定义 1 中，设 $\{X_t, t \in Z\}$ 是零均值平稳序列，对任意 t，满足线性差分方程

$$X_t - \varphi_1 X_{t-1} - \varphi_1 X_{t-1} - \cdots - \varphi_p X_{t-p} = Z_t + \theta_1 \varepsilon_{t-1} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

其中， $\{Z_t\} \sim WN(0, \sigma^2)$， p 阶自回归多项式

$$\varphi(z) = 1 - \varphi_1 z - \cdots \varphi_p z^p$$

与 q 阶滑动平均多项式

$$\theta(z) = 1 - \theta_1 z - \cdots \theta_p z^p$$

无公共因子 $\varphi_p \neq 0, \theta_p \neq 0$，则称 $\{X_t, t \in Z\}$ 为 p 阶自回归、q 阶滑动平均序列，简称为 $ARIMA(p,q)$ 模型。 p， q 分别称为自回归阶数和滑动平均阶数，实参数 $\varphi_1, \cdots, \varphi_p$ 为自回归参数，实参数 $\theta_1, \cdots, \theta_p$ 为滑动平均参数。

如果记

$$\varphi(B) = 1 - \varphi_1 B - \cdots - \varphi_p B \quad \theta(B) = 1 - \theta_1 B - \cdots - \theta_p B$$

其中， B 是延迟算子，则 式可简记

$$\varphi(B) X_i = \theta(B) Z_i \quad i \in Z$$

显然， $ARIMA(p,0)$ 模型就是 $AR(p)$ 模型，而 p $ARMA(0,q)$ 模型就是 MA（q）模型。这样显然表示一般模型有 $p+q$ 个参数要估计，看起来很繁琐，但利用计算机软件则是常规运算，并不复杂。

定义 3：在定义 2 中，如果 $\theta(z) \equiv 1$，则称满足线性差分方程

$$\varphi(B) X_i = Z_i \quad i \in Z$$

零均值平稳序列 $\{X_t\}$ 为 p 阶自回归序列，简记 $AR(p)$ 序列，称 $\{X_t, t \in Z\}$ 满足 $AR(p)$ 型.

定义 4：在定义 3 中，如果 $\varphi(z)\equiv 1$，则称满足线性差分方程

$$X_t=\theta(B)Z_t$$

的零均值平稳序列 $\{X_t\}$ 为 q 阶滑动平均序列，简记为 $MA(q)$ 序列，这时称 $\{X_t,t\in Z\}$ 满足 $MA(q)$ 模型。

4.1.1 自回归移动平均模型

在全球流感监控组织所提供的 2007 年 3 月至 2009 年 1 月全球流感季节性分布内容与种类的数据中，包含世界各大洲人口主要分布国家在全年中不同月份不同病毒的爆发情况。病毒的爆发程度分为四个等级：偶然性爆发、当地爆发、区域性爆发和大范围爆发，其严重性与传播性以指数型性增长，我们近似认为，爆发程度每上升一级，其扩散权重因子就增加一倍，即为 2 的指数幂的函数关系。

$$V_{kij}=2^n(n=0,1,2,3,4)$$

在相同的地理环境、医疗条件下，人口的数量对于流感扩散影响较大，在考虑流感病毒的扩散因子时，在考虑人口比重因子：

$$D_{kij}=\frac{M_i}{M}$$

对一区域病毒 j 扩散权重数据计算为：

$$O_{kj}=\sum_i V_{kij}*D_{kij}$$

对世界范围内第 j 种病毒扩散权重：

$$W_j=\sum_i V_{kij}*D_{kij}*P_k$$

将所有数据进行处理后得到了流感病毒中主要分类 $H1$、$H3$、B、A 随时间的序列分布，一组依赖于时间 t 的随机变量，而这组随机变量具有自相关性表征预测对象发展的延续性。以下引入 ARIMA(Autoregressive Integrated Moving Average Model)，即自回归移动平均模型，通过从时间序列的过去值及现在值可以预测其未来值。

应用以上模型，对数据分析处理有如下三方面：1.数据平稳化预处理；2.模型的识别定阶与模型参数估计；3.模型的诊断检验。

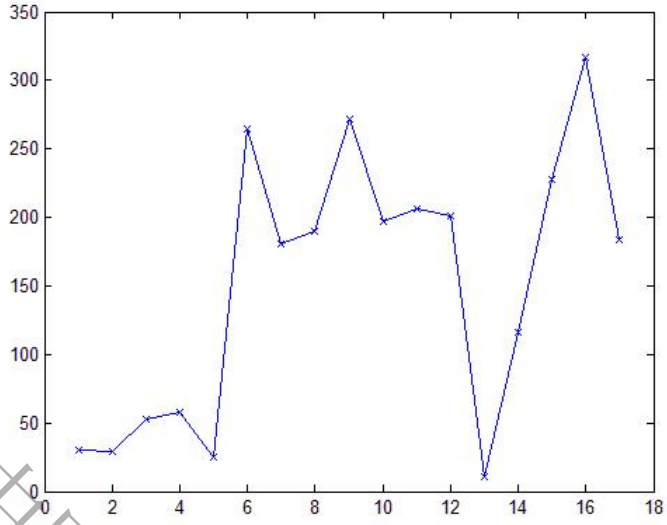
4.1. 1 数据平稳化预处理

平稳时间序列可以看作一种线性转换装置，它将白噪声(white noise)信号转换为所描述的时间序列。白噪声序列的定义为:若随机序列 $\{y_t\}$ 由互不相关的随机变量构成，即对所有的 $s\neq t,Cov(y_s,y_t)=0$，则称其为白噪声序列。时间序列的平稳性可通过其数据图和自相关函数来判断。如果一个序列的平均值和方差始终为常数，则称它为平稳的。如果数据图呈现线性或非线性趋势，则时间序列是不平稳的。

将亚洲 2006 年 11 月至 2008 年 2 月的亚洲 $H1$ 亚型流感病毒的权值数据导入 MATLAB 中，利用其工具箱中函数进行时间原始序列拟合，得到如下的原始时间序列图（其中 x

轴为月份，y 轴为 $H1$ 的权值：

图 4-1 原始时间序列图

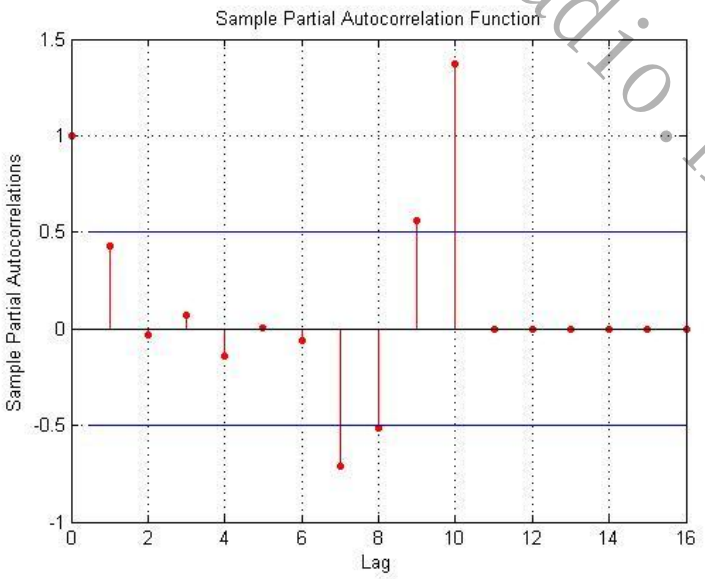


同时可以根据自相关函数和偏自相关函数得到自相关分析图（*ACF*）和偏自相关分析图（*PACF*）. 其中 *ACF* 图显示出典型的短期相关性，而 *PACF* 图则显示出函数阶数的截尾性，其能将时间序列的各阶滞后的自相关和偏自相关函数值及其在一定得置信水平下的置信区间直观地展现出来，较函数值和统计检验更为直观和方便。根据 *AIC* 准则（最小信息准则），未知参数越多，参数估计的精度就越差。分析该图中由自相关函数可得，其前面几个少数的数下降为零，其时间序列性平稳。

在以下的自相关分析图（*ACF*）和偏自相关分析图（*PACF*）中可以看出，相关函数值都分布在 95%的置信区间中，可将该模型应用于解决该问题中。

图 4-2 自相关分析图（*ACF*）

图 4-3 偏自相关分析图（*PACF*）



4.1. 2 模型的识别定阶与模型参数估计

根据 *ACF* 图的 4 阶截尾性，我们可以构造相关的线性相关函数为：

$$X_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \varphi_3 x_{t-3} + \varepsilon_t$$

则残差尾：

$$\varepsilon_t = X_t - \varphi_1 x_{t-1} - \varphi_2 x_{t-2} - \varphi_3 x_{t-3}$$

最终相关函数为：

$$X_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \varphi_3 x_{t-3} + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \phi_3 \varepsilon_{t-3} + \phi_4 \varepsilon_{t-4}$$

根据数据导入 *MATLAB* 中运行可得该亚型独立扰动值（运行程序与结果请见附录）：

$$\varepsilon_t = 0.1531\varepsilon_{t-1} - 0.0295\varepsilon_{t-2} - 0.2206\varepsilon_{t-3} - 0.4025\varepsilon_{t-4}$$

4.1.3 模型的诊断检验

模型的适合程度，需要对其拟合优度进行检验，典型方法是对观测值和弄醒的拟合值的残差进行分析。如果残差序列不是白噪声序列，则说明还有信息包含在相关的残差序列中未被提取，则此时可对残差拟合更复杂的模型以达到更适合的情况。

ARIMA 模型的诊断检验，即模型的残差序列 ε_t 的独立性检验，一般采用残差序列的卡方检验，公式为

$$\chi_m^2 = \frac{n(n+2)\sum_{k=1}^{m} r_k^2}{n-k}$$

其中

$$r_k^2 = \frac{\sum_{t=1}^{n-k} \varepsilon_t \varepsilon_{t+k}}{\sum_{t=1}^{n} \varepsilon_t}$$

如果残差序列不是白噪声序列，则需重新建立模型，重复上述步骤，直到残差序列是白噪声序列为止。我们利用亚洲 2007.3-2008.2 有关 *H*1 数据代入导入 *SPSS*(13.0) 中，可得到其残差诊断的数据如下：

表 4-1 *ARIMA* 模型中的残差诊断

Residual Diagnostics

Number of Residuals	15
Number of Parameters	2
Residual df	11
Adjusted Residual Sum of Squares	38226.659
Residual Sum of Squares	52864.327
Residual Variance	3073.123
Model Std. Error	55.436
Log-Likelihood	-80.232

Akaike's Information Criterion (AIC)	**36.464**
Schwarz's Bayesian Criterion (BIC)	71.296

AIC 准则目的是判别预测目标的发展过程与随机过程的接近性。因为只有当样本量足够大时，样本的自相关函数才非常接近母体的自相关函数。本问题中 $AIC = 36.464$，其数值较小，同时此时的自相关分析图（*ACF*）和偏自相关分析图（*PACF*）如下，可见其相关函数值都分布在 95%的置信区间中，可见模型有一定得可行性。

图 4.4 自相关分析图（*ACF*）

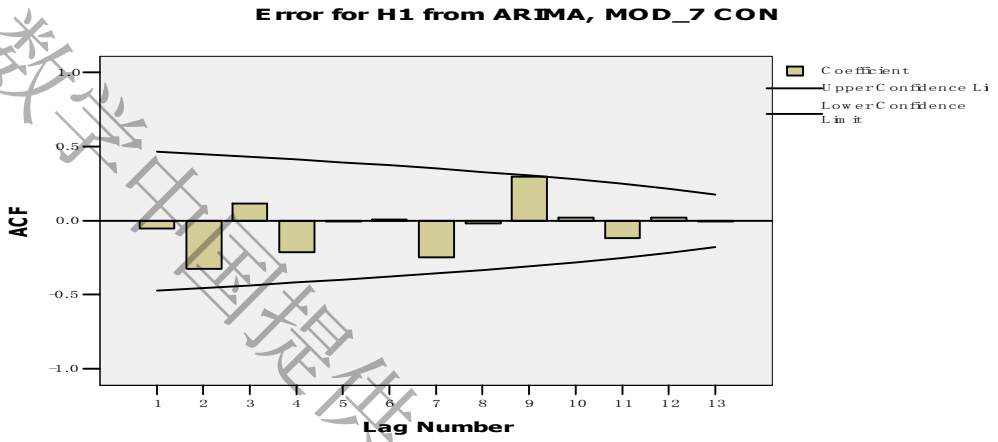
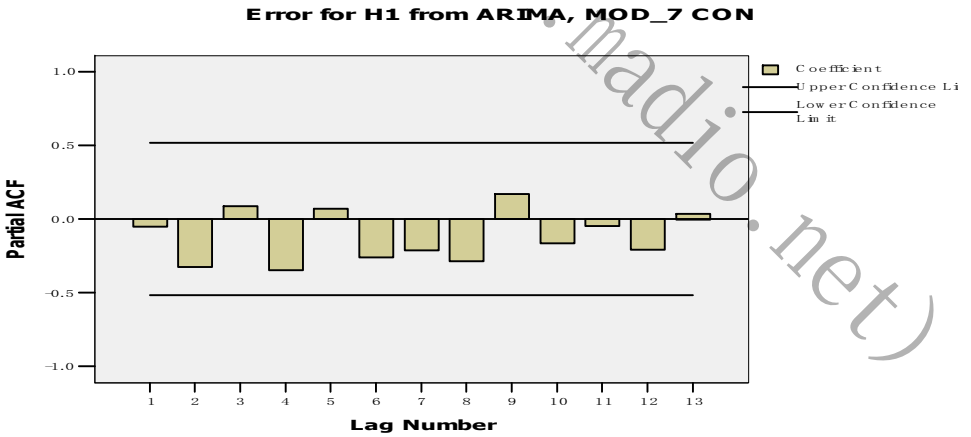


图 4-5 偏自相关分析图（*PACF*）



4.1.2 问题一模型的结果

4.1.2.1 模型结果的得出

根据上述数据分析，光滑曲线散点图对 2007.3-2008.2 时间段的描述，根据 *ARIMA* 模型中所求得的独立扰动值，可以计算出每个地区每种病毒在当地流行的比重，双联装装疫苗的成分则取比重较大的两种亚型作为疫苗成分，用其来预测 2008.10-2009.1 的数据，与 2008.10-2009.1 的实际数据相比较，在误差允许范围之内，模型具有可行性。

模型预测的三联装改进为二联装在全球的七大区域方案如下：

表 4-2 全球冬季疫苗改装预测

预测	区域	双联装中所包含的亚型
北半球 (2008.10-2009.01)	北非	$H1,B$
	北美洲	$H1,A$
	亚洲	$H3,B$
	欧洲	$H1,H3$
南半球 （2008.04-2008.07）	撒哈拉以南非洲	$H1,B$
	南美洲	$H1,B$
	大洋洲	$H3,B$

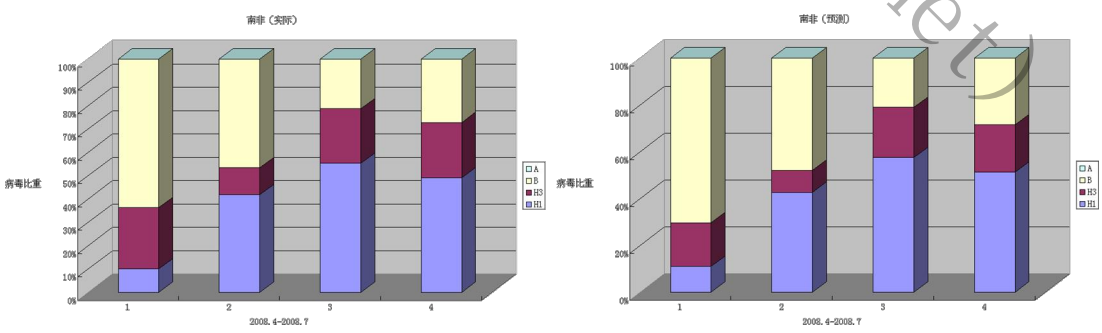
4.1.2.2 模型结果的诊断检验

根据以上方法，在南北半球各选取典型区域进行一下诊断检验，经数据验证可得该序列为白噪声序列，且据 2008.10-2009.1 的南非实际流感病毒的数据，我们可以得到如下图 4-4 的柱状图，其中青色表示 A，黄色表示，B 紫红色表示 $H3$，蓝色表示 $H1$，在 Y 轴上对应的比例便是该病毒在四种病毒中所占的比重。我们通过 2007.3-2008.2 的数据作为既得的原始数据，而通过模型求解可以求得 2008.10-2009.1 的数据，从而检验模型的正确性。

根据以上分析数据，算法设计，公式假设，函数导出，可得到如下右图 4-5 的柱状图。在图中可清晰看到，两者数据基本一致，模型具有可行性，可用于预测 2009 年冬季流感高峰期双联装疫苗的选择使用。

南半球

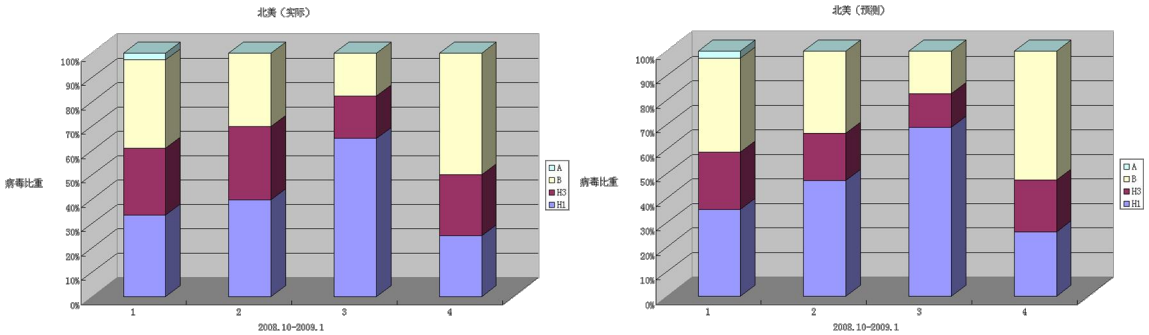
图 4-6　南非（实际）与 南非（预测）



附：

北半球

图 4-7　欧洲（实际）与　欧洲（预测）



4.2 问题二的模型建立与求解

4.2.1 问题二模型的建立

对于评估和预测下一年的优势毒株，对 2001-2008 年这 7 年的数据进行分析，并根据权值计算公式计算出各个区域内病毒亚型的权值，再将它乘以每个区域人口占世界总人口的百分比得到全球范围内的各病毒亚型所占百分比。

对世界范围内第 j 种病毒扩散权重：

$$W_j = \sum_i V_{kij} * D_{kij} * P_k$$

（其中 k 为区域编号，i 为国家编号，j 为病毒编号，）

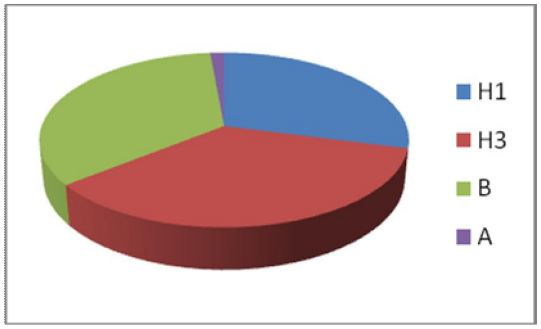
以下利用聚类分析的方法对于既得的病毒的权重进行分析。

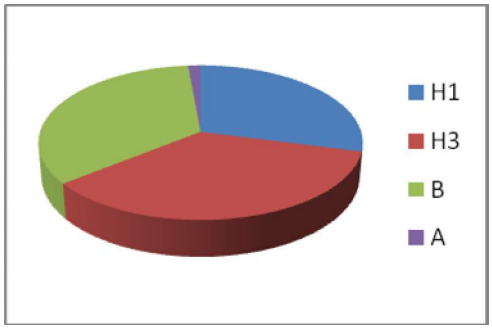
聚类分析是指在所选变量的基础上对样本数据进行分类，分类结果是各个变量综合计量的结果，以个体间的距离来度量个体间的“亲疏程度”。对于个体间的“亲疏程度”的度量一般有两个角度：一是个体间的相似程度，在问题二中，我们将国家作为个体，而在地理区域上相邻的国家在人种、医疗水平、经济状况等因素上相似，我们可以对于这些国家进行聚类，从而简化问题；二是个体间的差异程度，而个体间的差异程度通常用距离来测度。距离是指将每个样品看成是 m 个变量在 m 维空间中的一个点，然后在该空间中所定义，距离越近则紧密程度越高。通过对记得的数据进行聚类分析，我们可以得到下一年威胁较大的病毒所处位置的最大可能性，以便于专家的研究和疫苗的生产与投放。

4.2.2 问题二模型的求解

由数据可以发现对于流行毒株种类的变更主要出现在流行季节的后期，所以当其用于估计与预测下一年的优势毒株时，需要重点关注流感高发区域——北美洲和亚洲，重点时间段（北半球当年的 12 月与下一年 1，2 月）的毒株分离情况，对其中的毒株加以研究就可以准确预测出下一年流行季节的优势毒株而确定出下一年的疫苗成份。

图 4-8 北美洲与亚洲 2007.3-2008.2 病毒亚型比例





北美洲 亚洲

4.3 问题三的模型与求解

4.3.1 聚类分析原理与模型

4.3.1.1 聚类分析原理

聚类分析（cluster analysis）,是多元统计学中应用极为广泛的一类重要方法。为解决以下问题，我们引入聚类分析中的分层聚类的分析方法。应用分层聚类Q型聚类对样本进行聚类，即将具有相似的样本聚集在一起，使差异大的样本分离开来。

就分层聚类法的聚类方式而言，可以分为凝聚方式聚类和分解方式聚类。在解决该问题中，我们采用的是凝聚方式聚类，其过程是：首先将n个国家看出n类（一类包括一个国家），然后将性质最接近的两类合并成一类(性质接近一般指距离接近)，从而得到$n-1$类；接着从中找出最接近的来那个两类加以合并变成$n-2$类。重复这一过程，直到所有的国家都在一类当中。可见，在凝聚方式聚类过程中，随着聚类的进行，类内的“亲密”程度在逐渐降低，对n个国家通过$n-1$步可以凝聚成一大类。

在分层聚类中涉及对于个体间“亲疏程度”的度量和对于小类间“亲密程度”的度量。个体间“亲密程度”的度量，则可以先定义国家与小类、小类与小类之间的距离，距离小的关系比较密切，距离大的关系比较弱。

4.3.1.2 聚类分析数学模型

据世界卫生组织报告中可得，世界上共有193个与世界卫生组织建立合作的国家，而某些国家在国家经济条件，医疗水平等方面有一定的相似性，可以根据其相关参数的相近性对国家进行聚类。我们假设第i个国家卫生支出占GDP的百分比为α_i，对其进行分层聚类， 在193个国家中选取具有代表性的20个国家的α_i的值导入SPSS软件中，可以得到分层聚类中的凝聚状态表如下，此表中给出了系统聚类过程中的聚类信息，即每一个类被合并的类、被合并类间的类间距离以及最终的类水平。

表 4-3 分层聚类中的凝聚状态表

Agglomeration Schedule

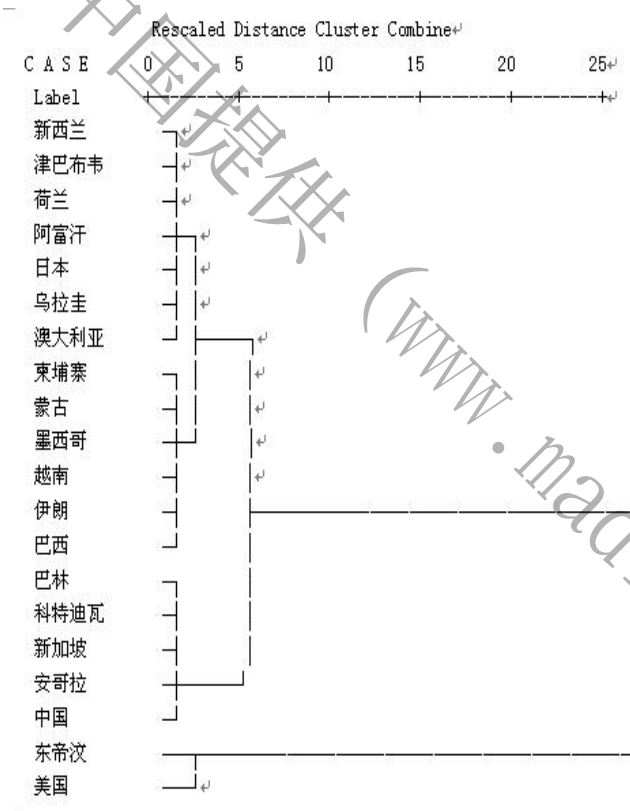
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
1	12	19	.000	0	0	4
2	9	16	.000	0	0	8
3	8	20	.010	0	0	10
4	11	12	.010	0	1	5
5	1	11	.020	0	4	13
6	4	18	.040	0	0	9
7	6	10	.040	0	0	14
8	7	9	.040	0	2	11
9	4	13	.170	6	0	12
10	3	8	.305	0	3	13
11	5	7	.703	0	8	14
12	2	4	.977	0	9	15

13	1	3	1.008	5	10	17
14	5	6	1.302	11	7	17
15	2	17	1.832	12	0	18
16	14	15	5.760	0	0	19
17	1	5	6.225	13	14	18
18	1	2	19.867	17	15	19
19	1	14	103.914	18	16	0

在表 4-3 中，第一列（Stage）表示聚类分析的步数；第二列，第三列（Cluster Combined）表示这一步聚类中哪两个国家或小类聚成一类；第四列（Coefficients）是国家距离或小类聚类；第五列和第六列（Stage Cluster First Appear）表示这一步聚类中参与聚类的是国家还是小类，0 表示国家，非 0 表示由第 n 步聚类生成的小类参与本步聚类；第七列（Next Stage）表示本步聚类的结果将在以下第几步中用到。

同时生成分层聚类的树状图如下：

图 4-9 分层聚类树状图



在图 4-9 中，树形图以躺倒树的形式展现聚类分析中的每一次合并的情况。SPSS 自动将各类间的距离映射在 0-25 间，并将聚类过程近似地表示在图上，在图上可以很清晰地看到聚类的过程。

将上述步骤应用到 193 个国家上，就可以得到其相应的凝聚状态表、冰柱图和树状图。最终可以将 193 个国家划分成 8 类，具体划分情况如下表所示：

表 4-4 193 个国家经分层聚类后类型一览表

类型一	阿尔及利亚	类型一	萨摩亚	类型四	安道尔	类型八	阿尔巴尼亚
	安提瓜巴布达		沙特阿拉伯		澳大利亚		巴巴多斯
	亚美尼亚		塞拉利昂		巴哈马		白俄罗斯

第二届数学中国数学建模网络挑战赛 #1036

	阿塞拜疆		新加坡		博茨瓦纳		玻利维亚
	巴林		所罗门群岛		巴西		布隆迪
	孟加拉国		斯里兰卡		保加利亚		柬埔寨
	伯利兹		苏丹		布隆迪		塞浦路斯
	贝宁		阿拉伯叙利亚		哥伦比亚		捷克共和国
	不丹		塔吉克斯坦		哥斯达黎加		刚果民主
	喀麦隆		泰国		克罗地亚		吉布提
	佛得角		汤加		古巴		多米尼克
	中非共和国		特立尼达多巴哥		芬兰		多米尼加
	乍得		突尼斯		格鲁吉亚		埃及
	智利		土耳其		海地		萨尔瓦多
	中国		土库曼斯坦		匈牙利		格林纳达
	科摩罗		乌兹别克斯坦		爱尔兰		几内亚
	库克群岛		瓦努阿图		以色列		几内亚比绍
	科特迪瓦		委内瑞拉		日本		圭亚那
	朝鲜		也门		黎巴嫩		洪都拉斯
	厄瓜多尔		阿富汗		卢森堡		伊朗
	厄立特里亚		阿根廷		马尔代夫		吉尔吉斯坦
	爱沙尼亚		奥地利		马耳他		拉脱维亚
	埃塞俄比亚		比利时		黑山		莱索托
	斐济		波斯尼亚		挪威		立陶宛
	加蓬		加拿大		巴拿马		马里
	冈比亚		丹麦		巴拉圭		墨西哥
	加纳		法国		圣马力诺		蒙古
	危地马拉		德国		塞尔维亚		尼日尔
	印度		希腊		斯洛伐克		波兰
	伊拉克		冰岛		斯洛文尼亚		大韩民国
	牙买加	类型二	意大利		南非		圣基茨
	哈萨克斯坦		约旦		西班牙		圣卢西亚
	肯尼亚		荷兰		马其顿共和国		圣文森特
	老挝		新西兰		乌干达		圣多美
	利比里亚		尼加拉瓜		北爱尔兰		塞内加尔
	马达加斯加		帕劳		乌拉圭		塞舌尔
	马来西亚		葡萄牙		安哥拉		苏里南
	毛里求斯		摩尔多瓦		文莱达鲁萨兰		斯威士兰
	摩纳哥		卢旺达		刚果		多哥
	摩洛哥		瑞典		赤道几内亚		乌克兰
	莫桑比克		瑞士	类型五	印度尼西亚		坦桑尼亚
	纳米比亚		图瓦卢		科威特		越南
	尼泊尔		津巴布韦		阿拉伯利比亚		赞比亚
	尼日利亚	类型三	基里巴斯		毛里塔尼亚		
	巴布亚		马拉维		缅甸		
	秘鲁		马绍尔群岛		阿曼		

	菲律宾		密克罗尼西亚		巴基斯坦
	卡塔尔		瑙鲁		阿联酋
	罗马尼亚		纽埃	类别六	索马里
	俄罗斯联邦		美利坚合众国	类别七	东帝汶

4.3.2 国家优先级系数 γ 的得出

联合国的经费投入到不同国家建立实验站与哨点医院，经费投放的选择与该国的哨点原有的哨点医院数，医生密度等因素有关。医生数量则数该医院规模的重要反应条件之一。我们假设第 i 国家哨点医院医生占为医生总数的比值为 β_i，则可以定义：

$$\beta_i = \frac{S_{mi}}{S_m}$$

其中 S_{mi} 表示第 i 个国家中哨点医院的医生数， S_m 表示该国医院的医生总数。

则该国哨点医院的医生密度 δ_i 可表示为：

$$\delta_i = \rho_i * \beta_i$$

其中 ρ_i 为每 10000 人的医生密度。

对哨点医院的医生密度 δ_i 数据进行标准化，即通过检索工具确定哨点医院的医生密度的最大值 $\delta_{\max}$ 与最小值 $\delta_{\min}$，以此确定数据变化的基本范围，确定其数字化值，通过下式确定：

$$l = \delta_{\max} - \delta_{\min}$$

然后将其统一定义域范围到（0-1）间，可完成其标准化后哨点医院的医生密度 ϕ 为：

$$\phi = \frac{(\delta_i - \delta_{\min})}{l}$$

我们可以将国家的优先级系数 γ 定义为：

$$\gamma = 1 - \phi$$

则系数 γ 越大，则该国家的优先级就越高，需要得要联合国的援助就越紧急。

五、模型的评价

模型一的目的在于降低疫苗的制造成本，将三联装的疫苗改装成双联装，如此设想主要在于两方面考虑：一是对于某些地区，盛行并造成大规模影响的病毒只有两种，三联装疫苗造成经济上的浪费；二是对于某些病毒，其生命周期短，环境适应力低，疫苗对其作用不大。以下引进覆盖率的概念来评价改装的价值：覆盖率表示疫苗抵抗病毒占总病毒的百分比。表中三联装疫苗的覆盖率与双联装疫苗的覆盖率相比较，差值均在33.3%以下，可见改装后疫苗与原疫苗相比有一定的进步。

表 4-2 疫苗覆盖率比较

	月份	原三联装疫苗对病毒亚型的覆盖率	改装后双联装疫苗对病毒亚型的覆盖率
撒哈拉以南非洲	4	100.00%	81.31%
	5	100.00%	90.53%
	6	100.00%	78.53%
	7	100.00%	79.69%
北美洲	10	99.02%	75.95%
	11	100.00%	80.68%
	12	100.00%	86.28%
	1	100.00%	78.90%
欧洲	10	89.78%	73.81%
	11	96.42%	71.45%
	12	97.16%	82.13%
	1	98.23%	84.50%

六、模型的推广

6.1 疫苗生产预测

根据上述模型的分析与研究，我们可以看到在不同区域不同病毒亚型的流行情况，根据这些情况，我们运用问题一模型中的权值确定该地区该种亚型的权值，权值越大，证明该种亚种在该地区的爆发与流行的可能性就越大，因此在生产疫苗时就因在控制成本的前提下选用针对该种病毒亚型的疫苗。

通过 2007.3-2009.1 月的所有数据，我们可以预测出下一年度流感爆发的高峰期的病毒亚型比重情况，为疫苗的研制提供有效的数据参考。北半球与南半球数据分别如下：

表 5-1 北半球数据

区域	病毒亚型	时间							
		2009.10		2009.11		2009.12		2010.01	
		权值	百分比	权值	百分比	权值	百分比	权值	百分比
北非	H1	127.84	100%	127.84	83%	127.84	51%	127.03	35%
	H3	0.00	0%	26.38	17%	54.24	22%	36.58	10%
	B	0.00	0%	0.00	0%	68.86	27%	200.00	55%
	A	0.00	0%	0.00	0%	0.00	0%	0.00	0%
北美	H1	158.64	30%	200.00	37%	561.55	65%	318.27	41%

洲	H3	162.54	30%	182.26	34%	169.24	20%	156.31	20%
	B	200.00	37%	151.85	28%	132.16	15%	303.70	39%
	A	15.29	3%	0.00	0%	0.00	0%	0.00	0%
亚洲	H1	129.24	39%	203.45	40%	287.64	45%	193.84	38%
	H3	83.56	25%	109.04	22%	142.24	22%	179.54	36%
	B	115.37	35%	191.23	38%	204.89	32%	132.34	26%
	A	0.00	0%	0.00	0%	0.00	0%	0.00	0%
欧洲	H1	83.21	28%	111.29	24%	216.56	25%	239.88	20%
	H3	97.34	32%	186.51	40%	518.12	59%	798.23	66%
	B	107.22	36%	157.34	34%	142.56	16%	178.24	15%
	A	13.18	4%	5.68	1%	2.16	0%	0.00	0%

根据上述数据，我们可以得到北半球冬季二联装疫苗成分的预测：

表 5-2 北半球预测

预测	区域	双连装中所包含的亚型
北半球 (2008.10-2009.01)	北非	$H1, B$
	北美洲	$H1, H3$
	亚洲	$H1, B$
	欧洲	$H3, B$

同理，可得到有关南半球的情况：

表 5-2 南半球数据

区域	病毒亚型	时间							
		2009.04		2009.05		2009.06		2009.07	
		权值	百分比	权值	百分比	权值	百分比	权值	百分比
撒哈拉以南非洲	H1	73.24	29%	147.35	39%	322.79	55%	339.16	47%
	H3	42.78	17%	68.29	18%	136.80	23%	185.59	25%
	B	132.60	53%	158.22	42%	124.49	21%	204.58	28%
	A	0.00	0%	0.00	0%	0.00	0%	0.00	0%
南美洲	H1	402.36	40%	1232.37	62%	617.42	47%	686.54	34%
	H3	19.76	2%	19.76	1%	0.00	0%	0.00	0%
	B	318.37	32%	438.31	22%	586.49	45%	896.06	45%
	A	258.60	26%	312.35	16%	107.98	8%	410.48	21%
大洋洲	H1	0.00	0%	0.00	0%	165.74	14%	165.74	9%
	H3	0.00	0%	200.00	50%	583.68	48%	937.04	51%
	B	165.74	100%	200.00	50%	468.52	38%	732.53	40%
	A	0.00	0%	0.00	0%	0.00	0%	0.00	0%

根据上述数据，我们可以得到南半球冬季二联装疫苗成分的预测：

南半球 （2008.04-2008.07）	撒哈拉以南非洲	$H1,B$
	南美洲	$H1,B$
	大洋洲	$H3,B$

七、参考文献

[1] 姜启源、谢金星、叶　俊，《数学模型》，北京：高等教育出版社，2003.8
[2] 苏金明、阮沈勇，《MATLAB 实用教程》，北京：电子工业出版社，2008.2
[3] 欧春泉、邓卓辉、杨　琳、陈平雁，《用自回归模型预测流感样病例数的变化趋势》，《中国卫生统计》，第 24 卷第 6 期：569-571，2007.12
[4] 张善文、雷英杰、冯有前，《MATLAB 在时间序列分析中的应用》，西安：西安电子科技大学，2007.4
[5] 张树京、齐立心，《时间序列分析简明教程》，北京：清华大学出版社，2003.9
[6] 小田切孝人，《流感病毒感染性疾病和相关疫苗的临床应用专辑》，www.cpvip.com，2008 年 4 月 25 日 23：24
[7] 漆莉，《重庆市 2004-2006 年流感流行特征分析专辑》，www.cnki.net.2008.4.26.11:23
[8] 朱大方，《江苏省流感症状监测方法研究》，www.cnki.net. 2008.4.27.1:03
[9] 刘大海、李宁、晁阳，《SPSS 15.0 统计分析》，北京：清华大学出版社，2008.5

八、附录

EDIT1.m

```
%对原始时间序列进行拟合的MATLAB源码
N=17;
n=3;
m=N-n;
X=[];
for i=1:m
for j=1:n
X(i,j)=H1(n+i-j);
end
end
HH=H1(n+1:N)';
[b,bint,r,rint,stats]=regress(HH,X)
autocorr(r)
```

运行结果如下：

```
b =
```

```
    0.6303
    0.0632
    0.3010


bint =

   -0.0311    1.2918
   -0.7803    0.9067
   -0.4370    1.0390


r =

   13.6719
  -23.9505
  228.8948
   -5.1539
   51.7806
   61.4986
  -40.7646
    7.5744
  -22.6969
 -188.4179
   34.1912
   93.4833
  162.1308
  -64.9654


rint =

 -243.6473  270.9911
 -280.6198  232.7188
   28.1810  429.6087
 -204.9463  194.6385
 -136.6680  240.2292
 -163.1792  286.1764
 -277.1090  195.5799
 -222.5150  237.6639
 -250.1736  204.7798
 -392.6943   15.8585
 -167.0218  235.4042
 -104.8699  291.8365
```

```
  -36.1859   360.4475
 -283.5908   153.6601


stats =

  1.0e+004 *

   -0.0000    0.0003    0.0000    1.2527
```

EDIT2.m

```
%对残差序列进行拟合的 MATLAB 源码
N=12;
n=4;
m=N-n;
X=[];
for i=1:m
for j=1:n
X(i,j)=ri(n+i-j);
end
end
R=ri(n+1:N);
[b,bint,r,rint,stats]=regress(R,X)
autocorr(r)
```

运行结果如下：

```
b =

    0.1531
   -0.0295
    0.2206
   -0.4025


bint =

   -1.3077    1.6138
   -0.9451    0.8862
   -0.9386    1.3798
   -1.6652    0.8601


r =
```

```
   70.1048
   -6.7173
   44.6258
    2.1288
  -17.7816
 -150.9717
   44.2809
   90.7515


rint =

 -112.0308  252.2404
 -126.8895  113.4549
  -47.3868  136.6384
 -296.4633  300.7208
 -310.6446  275.0813
 -331.6735   29.7301
  -35.8503  124.4121
 -109.9403  291.4433


stats =

  1.0e+004 *

    0.0000    0.0000    0.0001    1.0065
EIDIT3.m
%求解预测值的 MATLAB 源码
n=6;
 Y=[];
 for i=1:6
     x=[H1(i+1:i+3),r(i+1:i+4)];
     Y(i)=b*x';
 end
     Y
```


世界各国人口数量（万人）

Europe	66127	**Southern Africa**	15270
Austria	818	Ghana	1890
Belarus	1015	Kenya	3000
Belgium	1022	Madagascar	1635
Bulgaria	823	Mauritius	1300
Croatia	450	Senegal	900

Czech Republic	1028	South Africa	4305
Denmark	531	**Southern America**	25107
Finland	517	Argentina	3660
France	6019	Brazil	16390
Germany	8209	Chile	1510
Greece	1060	Costa Rica	360
Hungary	1009	France, French Guiana	15
Israel	710	France, Martinique	38
Italy	5734	Honduras	654
Latvia	229	Peru	2480
Luxembourg	43	**Northern America**	40000
Netherlands	1586	Canada	3057
Norway	445	Mexico	9630
Poland	3865	United States of America	27313
Portugal	998	**Asia**	272159
Romania	2252	China	130032
Russian Federation	14550	China (Province of taiwan)	47
Serbia	1015	China, Hong Kong RAS	685
Slovakia	530	India	102700
Slovenia	199	Iran	7005
Spain	3940	Japan	12678
Sweden	886	Mongolia	245
Switzerland	724	Philippines	7515
U. K	5920	Republic of Korea	4685
Northern Africa	10322	Singapore	389
Egypt	6598	Thailand	6178
Morocco	2780	**Oceania**	2499
Tunisia	944	Australia	2071
Uganda	2240	New Zealand	428