钩端螺旋体病流行因素的聚类和逐步回归分析

四川省涪陵地区卫生防疫站* 李优良 甘业光 刘 洪
四川省垫江县卫生防疫站 方廷甫

提要 通过对稻田型钩端螺旋体病严重流行区连续监测10年的22项流行病学资料分三类进 行聚类分析，选出5个典型变量进行逐步回归分析，计算出多元回归方程。该方程引入了鼠密度、主要带菌鼠种的带菌率、人群自然抗体GMT、8月降雨量4个因素。经验证该方程计算出的理论 年发病率与实际年发病率基本一致。

关键词 钩端螺旋体病 聚类分析 逐步回归分析

近年在钩体病研究中，应用单因素进行理论流行病学分析已有报道[1~2]。为了提高监测质量，探讨监测方法和预测指标的可靠性，本文采用聚类分析和逐步回归分析方法，对1980～1989年在垫江县双河乡稻田型钩体病重流行区连续监测10年的22项流行因素进行数理分析，现报告如下。

材料和方法

一、监测县钩体病概况：垫江县属浅丘地貌，水田旱地相间，稻田是旱地的3倍。8月下旬和9月上旬是水稻收割季节。9月上中旬是钩体病发病高峰期。自1958年报告疫情以来，共发生四次较大流行，年发病率40～60/10万，爆发流行年发病率高达280/10万。1963年从患者血中分离出黄疸出血群钩体，血清学证实为黄疸出血群、七日热群等5个血清群。1975年对该县进行系统调查；1980～1989年在双河乡进行系统监测，年发病为20/万左右；1987年爆发流行发病率高达98.67/万，均证实黑线姬鼠、大足鼠、四川短尾鼩、大家鼠为主要带菌鼠种。先后从病人血中分离钩体91株，78株为黄疸出血群。从传染源肾分离出钩体216株，193株为黄疸出血群。从稻田水中分离出2株黄疸出血群钩体。

二、资料来源和方法：

1.资料来源：疫情资料采用垫江县卫生防疫站1980～1989年钩体病资料，年发病率以/10万计算。气象资料由垫江县气象局提供。其余各项资料均由监测组现场获得。每年流行季节前按统一方案在双河乡进行监测。该县钩体病疫情报告的准确率为86.14%。分析前，对历年各种资料进行审核后再列入统计。

2.分析方法：依据聚类分析法[3]将22项流行因素进行分类分析，选出有代表性的流行因素作为自变量，以该县年发病率为因变量进行逐步回归分析[4]。计算工具用IBM-PC/XT微机进行运算。

结 果

一、分析因素：本项研究列入分析指标共22项，分别为鼠密度%（X_1）、主要带菌鼠种的带菌率%（X_2）、黑线姬鼠数量%（X_3）、四川短尾鼩数量%（X_4）、大足鼠数量%（X_5）、大家鼠数量%（X_6）、黑线姬鼠带菌率%（X_7）、四川短尾鼩带菌率%（X_8）、大足鼠带菌率%（X_9）、大家鼠带菌率%（X_{10}）、人群抗体阳性率%（X_{11}）、

---

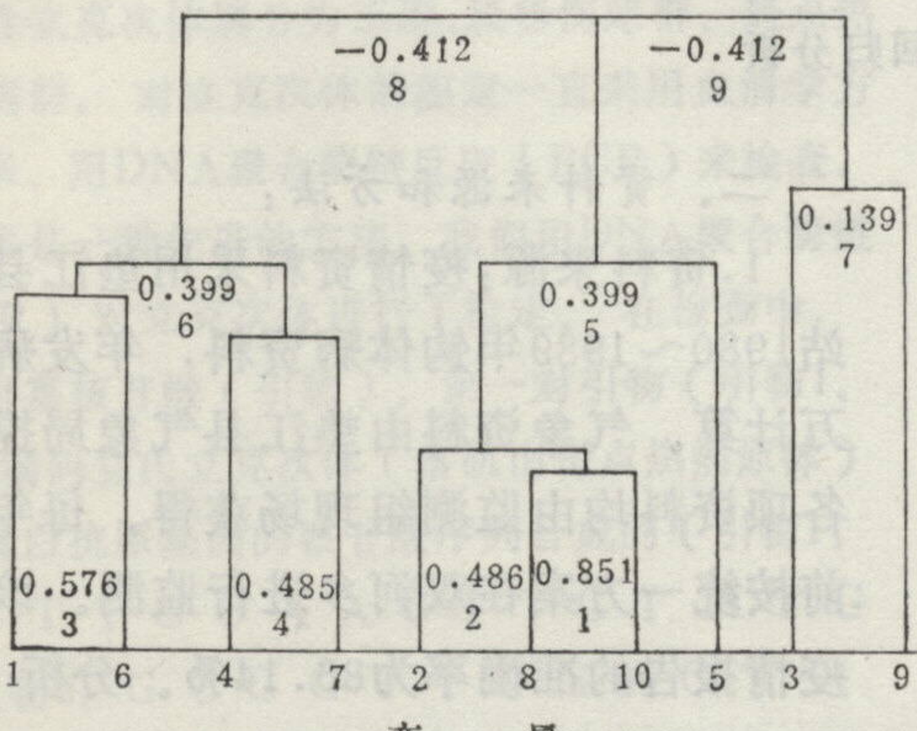
* 邮政编码 648000

人群自然抗体GMT（X_{12}）、阳性抗体 GMT（X_{13}）、年降雨量mm（X_{14}）、7月降雨量mm（X_{15}）、8月降雨量mm（X_{16}）、年降雨日数≥0.1mm/日（X_{17}）、7月降雨日数≥0.1mm/日（X_{18}）、8月降雨日数≥0.1mm/日（X_{19}）、年日照小时S（X_{20}）、7月日照小时S（X_{21}）、8月日照小时S（X_{22}）。

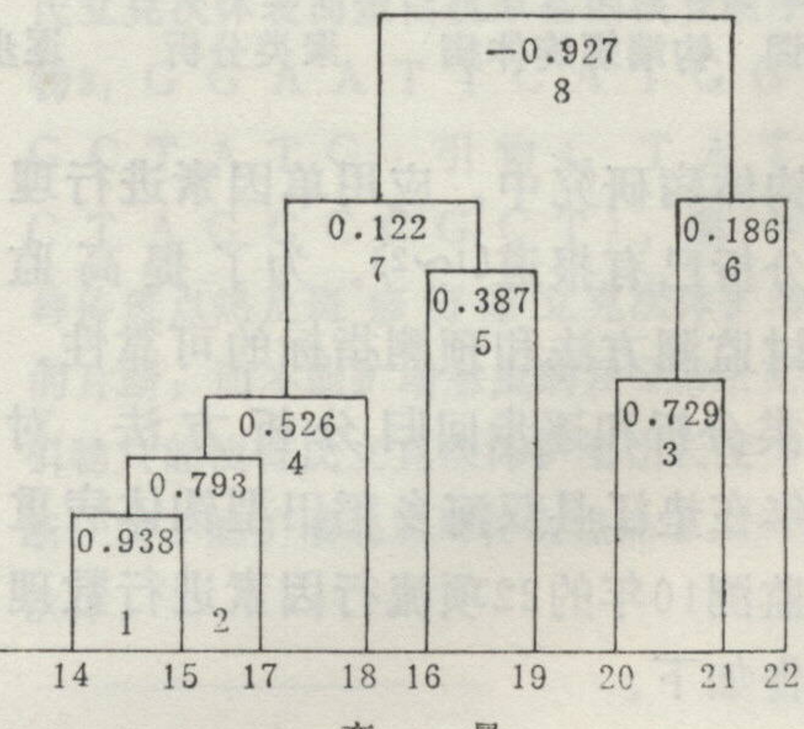
二、聚类分析：22项流行因素分属三个部分，即传染源、抗体、气象因素。传染源因素有X_1、X_2、X_3、X_4、X_5、X_6、X_7、X_8、X_9、X_{10}；抗体因素有X_{11}、X_{12}、X_{13}；气象因素有X_{14}、X_{15}、X_{16}、X_{17}、X_{18}、X_{19}、X_{20}、X_{21}、X_{22}。对各部分因素分别用R型聚类法进行聚类分析，结果见聚类图1、2。

从聚类图1、2及钩体病流行病学材料分析，聚类图1传染源因素可分为三类，其中X_1、X_4、X_6、X_7为第一类；X_2、X_5、X_8、X_{10}为第二类，X_3、X_9为第三类，聚类图2气象因素也可分为三类，即X_{14}、X_{15}、X_{17}、X_{18}为第一类，X_{16}、X_{19}为第二类，X_{20}、X_{21}、X_{22}为第三类，聚类后按每类中R^2值最大者挑选典型变量。抗体部分的三个因素经计算选出变量X_{12}。

22项流行因素经聚类分析共选出典型变量



聚类图1



聚类图2

5个，即X_1，X_3，X_{10}，X_{12}，X_{16}。

三、回归分析：在聚类分析的基础上，从第一类变量中选出X_1、X_3、X_{10}，第二类变量中选出X_{12}，第三类变量中选出X_{16}作为典型变量。其中$X_1R^2=0.194$，$X_3R^2=0.139$，$X_{10}R^2=0.578$，$X_{12}R^2=0.431$，$X_{16}R^2=0.013$。

以选出典型变量X_1、X_3、X_{10}、X_{12}、X_{16}与年发病率（Y）进行逐步回归分析，计算出多元回归方程式：

$$\hat{y}=-19.2468-3.3517_1+3.9362X_{10}-2.7305X_{12}+0.706X_{16}$$

回归系数$b_1=-3.3517$，$b_2=3.9362$，$b_3=2.7305$，$b_4=0.706$，$R=0.9653$，$P<0.001$，$L_{回归}=33686.33$，$L_{剩余}=2297.365$，$F=17.0788$，$P<0.001$，剩余标准差=21.4353，该回归方程引入了鼠密度X_1，主要带菌鼠种的带菌率X_{10}，人群自然抗体的GMTX_{12}，8月降雨量X_{16}四个流行因素。

该回归方程在钩体病疫情监测中回代结果，详见附表。附表表明其理论年发病率与实际年发病率基本一致，显示了该方程选入的流行因素组合较理想，监测和预测较准确，较稳定。

讨　论

采用聚类分析与逐步回归分析的方法应用于钩体病的监测和疫情预测国内研究尚未见报道，陈清等[1]应用人群抗体水平与年发病率推算出了监测钩体病的数式；作者等[2]应用8月降雨量与年发病率进行相关回归分析建立了预测钩体病疫情的一元回归方程式，由于因素

附表 理论值与实际值对照

年份	实际值(y)	预测值($\hat{y}$)	$y-\hat{y}$	$(y-\hat{y})^2$
1980	197.91	184.29	13.62	185.50
1981	45.46	43.12	2.34	5.48
1982	52.36	31.36	21.00	441.00
1983	71.14	107.57	−36.43	1327.14
1984	24.12	28.82	−4.70	22.09
1985	20.31	20.60	−0.29	0.08
1986	51.69	39.91	11.78	138.77
1987	167.29	160.09	7.20	51.84
1988	57.67	68.05	−10.38	107.74
1989	26.36	30.51	−4.15	17.22

单一，效果不够理想。

我国稻田型钩体病流行区流行病学调查证实，长江流域及以南的流行形式主要为稻田型，其主要特点之一传染源主要是在田间活动的鼠类，特别是黑线姬鼠、大足鼠、四川短尾鼩、东方田鼠、黄毛鼠等。其密度、带菌率与发病关系密切，尤其是当地主要带菌鼠种的数量及带菌状况是本病流行的最基本的条件。疫水是钩体病传播的媒介，而雨量是构成疫水的重要自然因素，特别是稻收前期和稻收期的降雨量与流行关系极为密切。人群对钩体的免疫水平在钩体病流行中居主要地位。当地流行菌株的毒力状况也与流行有重要关系。鉴于此，钩体病的流行是多因素共同作用下发生的。

本文建立的四元回归方程式10年回代结果，理论年发病率与实际年发病率基本一致。从而显示该方程因素选择与组合较好，在疾病监测和钩体病的疫情预测较准确和稳定。同时也进一步表明采用聚类分析和逐步回归分析方法作为钩体病流行病学研究方法是科学的，实用的。

多元回归模型可以评价各流行因素对钩体病流行程度的影响水平。本文建立的四元回归方程式，引入了鼠密度X_1（传染源的数量），主要带菌鼠种的带菌率X_{10}，人群钩体病的免疫水平GMTX_{12}及8月降雨量X_{16}四个流行因素，显示了四个流行因素在钩体病流行的互为协同的作用，表现出在流行中内在联系的全过程。因此，将这四个流行因素作为钩体病监测的四个指标是科学的、可靠的，对于影响因素较多、流行过程复杂的钩体病，我们建立的数学模型和提出的四个监测指标还需进一步探讨。

（本文承蒙重庆医科大学卫生系统计教研室周燕荣副教授、中国预防医学科学院流行病学微生物学研究所钩体病研究室聂第楷研究员、时曼华、张哲夫副研究员指导、审阅，一并致谢）

Cluster and Multiple Regression Analysis of Leptospirosis Epidemic Factors *Li Yuliang, et al., The Sanitary and anti-epidemic Station of Fuling Prefecture, Sichuan Province*

Ten-year surveillance of Rice-field-typed leptospirosis was carried out in high-infected foci. Twenty two parameters of epidemic data were divided into 3 categories and cluster-analyzed. 5 typical variables were selected for multiple regression equation. The equation contained four factors, i.e. rodents density, carriage ratse of main animal hosts, GMT of population antibody against leptospira and quantity of rainfall in August. The expected morbidity of leptospirosis calculated by the equation were roughly identical to the real morbidity.

Key words Leptospirosis Cluster and Multiple regression analysis

参考文献

1. 陈清，等. 用人群抗体水平探索钩端螺旋体病监测数式. 中国人兽共患病杂志1986；5(2)：41.
2. 李优良，等. 四川垫江县钩端螺旋体病一个流行周期(1980～1987)的监测. 中华流行病学杂志1989；10(特刊3号)：256.
3. 杨树勤，等.《中国医药百科全书、医学统计学》，第一版上海科学技术出版社. 1985.
4. 金丕焕，等.《医用统计程序集》上海科学技术出版社，1986.

（1991年7月收稿. 1991年9月1日修回）