·监测·

最优回归子集法在达乌尔黄鼠疫源地风险分级中的应用

周晓磊　张博宇　丛显斌　李仲来　姚晓恒　鞠成　徐成　张贵军
段天一　陈磊　刘振才

【摘要】 **目的**　对达乌尔黄鼠疫源地动物鼠疫流行情况进行风险分级。**方法**　对内蒙古达乌尔黄鼠疫源地动物鼠疫流行总体数据7个监测指标(鼠密度、鼠体染蚤率、鼠体蚤指数、巢穴蚤染蚤率、巢穴蚤指数、洞干蚤染蚤率、洞干蚤指数)利用Matlab软件中最优回归子集法进行风险分级,采用指数平滑法预测2012年动物鼠疫流行的风险。按照检出鼠疫菌为流行($y=1$),未检出菌视为不流行($y=0$),将风险分为流行、高风险及不流行3级,若预报值$y>2/3$,预报为流行;若预报值$y<1/3$,预报为不流行;若$1/3\leq y\leq 2/3$,预报为高风险。**结果**　对风险分级采用实际数据进行拟合,当$y>2/3$时预报流行的拟合率均为100%;回归模型的回归因子≥4个时,$y<1/3$时预报流行的拟合率均为100%;$1/3\leq y\leq 2/3$时预报流行的拟合率约为50%。**结论**　风险分级预测结果表明2012年达乌尔黄鼠疫源地不会发生动物鼠疫流行,预测结果与实际情况相符(当年实际并未检出鼠疫菌)。

【关键词】 最优回归子集法;达乌尔黄鼠疫源地;风险分级;预测

Application of Best Subsets Regression on the risk classification for Spermophilus Dauricus Focus

Zhou Xiaolei[1], *Zhang Boyu*[2], *Cong Xianbin*[1], *Li Zhonglai*[2], *Yao Xiaoheng*[1], *Ju Cheng*[1], *Xu Cheng*[1], *Zhang Guijun*[1], *Duan Tianyi*[1], *Chen Lei*[1], *Liu Zhencai*[1]. 1 *Chinese Base for Control of Plague and Brucellosis, Chinese Center for Disease Control and Prevention, Baicheng* 137000, China; 2 *Beijing Normal University*

Corresponding author: *Cong Xianbin*, Email: cxb0805@sina.com

This work was supported by a grant from the Research Special Fund of Health Sector of China (No. 201202021).

【**Abstract**】 **Objective**　To study the risk classification of animal plague in Spermophilus Dauricus Focus, using the Best Subsets Regression (BSR) model. **Methods**　Matlab, BSR and exponential smoothing were employed to develop and evaluate a model for risk classification as well as to forecast plague epidemics at the Spermophilus Dauricus Focus. Data was based upon the Inner Mongolia surveillance programs. This model involved 7 risk factors, including density of *Spermophilus dauricus*, percentage of hosts infested, host flea index, percentage of nests infested, nest flea index, percentage of runways infested, and runway flea index. **Results**　Forecasting values of the classification model (CM) were calculated and grouped into 3 risk levels. Values that over 2/3 of the CM would indicate the existence of potential epidemics while those below 1/3 would indicate that there were no risk for epidemics but when values that were in between would indicate that there exist for high risk. Annually, during the observation period in the Inner Mongolia Spermophilus Dauricus Foci, the detection of *Yersinia pestis* gave a risk rating value of 1 which stood for existing epidemics, while nil detection rate generated a 'zero' value which representing the situation of non-epidemic. The overall plague epidemics forecasting surveillance programs in 2012 at the Spermophilus Dauricus Foci indicated that no active plague was obsered. When the forecasting values became over 2/3, combinations of all the risk factors would achieve the consistency rates of 100%. When the forecasting values were below 1/3, combinations of at least the first 4 factors could also achieve the consistency

DOI: 10.3760/cma.j.issn.0254-6450.2014.02.015

基金项目:卫生行业科研专项项目(201202021)

作者单位:137000 白城,中国疾病预防控制中心鼠疫布氏菌病预防控制基地(周晓磊、丛显斌、姚晓恒、鞠成、徐成、张贵军、段天一、陈磊、刘振才);北京师范大学(张博宇、李仲来)

通信作者:丛显斌, Email: cxb0805@sina.com

rates of 100%. However, when the forecasting values fell in between, combinations of at least the first 4 factors would achieve the consistency rates of around 50%. **Conclusion** Results from our study showed that plague would not be active to become epidemic, in 2012.

【**Key words**】 Best subsets regression; Spermophilus Dauricus Focus; Risk classification; Forecasting

达乌尔黄鼠鼠疫自然疫源地已确定达乌尔黄鼠为该疫源地的主要宿主，褐家鼠为次要宿主，其他染疫动物为偶然参与流行的动物宿主[1,2]。目前国内对风险分级研究仅见于某些传染病及突发事件，在鼠疫研究中尚未见报道。本研究目的是对达乌尔黄鼠疫源地进行风险分级并预测，为预警提供理论支持。

资料与方法

1. 资料来源：达乌尔黄鼠疫源地分布于内蒙古自治区及东北3省（吉林、辽宁和黑龙江省）。在近30年的动物鼠疫监测资料中，只有内蒙古自治区动物鼠疫流行数据符合建立风险分级的要求，且代表性强，本研究采用其总体数据代表达乌尔黄鼠疫源地的总体数据。

2. 动物鼠疫流行判定标准：从疫源地的主要宿主及其媒介或其他动物体内检出鼠疫菌，或虽未检出鼠疫菌，但在疫源地内采用间接血凝方法从主要宿主或牧犬血清中连续检测到鼠疫F1抗体，其阳性率或其中有一份以上的血凝滴度达到各类型疫源地的判定标准之一者，即判定当年在检测区内曾发生或正在发生动物鼠疫流行[3]。本研究选择鼠密度（X_1）、鼠体染蚤率（X_2）、鼠体蚤指数（X_3）、巢穴蚤染蚤率（X_4）、巢穴蚤指数（X_5）、洞干蚤染蚤率（X_6）、洞干蚤指数（X_7）7项监测指标。

3. 统计学分析：采用Excel和Matlab软件（美国MathWorks公司）将数值分析、矩阵计算可视化以及非线性动态系统建模和仿真，并集成于易于使用的视窗内，摆脱了传统非交互式程序设计语言的编辑模式。

结　果

1. 建立最优回归方程：当选取4项（X_1、X_2、X_3和X_4）及其以上指标时，预测动物鼠疫流行的结果基本相同。将流行疫点（即动物中检出鼠疫菌）的因变量取值为$y=1$，未流行点取值为$y=0$，筛选出的最优回归模型及其统计检验见表1、2。

2. 鼠疫风险分级：由最优回归方程按年份计算出因变量y值（表3）。可见除1个因子（指标）外，含2～7个因子的最优回归模型计算得到的y值均满足$P<0.01$，即鼠疫发生和不发生时y值的差异有统计学意义，并以此作为风险分级的依据。

根据疫情发生情况及实际经验，本研究将预报数据的风险分为三级。预报值$y>2/3$，预报为流行；$y<1/3$，预报为不流行；$1/3\leqslant y\leqslant 2/3$，预报为高风险。对回归模型利用实际观察数据进行拟合（表4），当$y>2/3$时预测鼠疫流行的符合率均为100%；回归模型的因子≥4个时，$y<1/3$则预测流行的符合率均为100%，$1/3\leqslant y\leqslant 2/3$时，预测流行的符合率约为50%。

3. 2012年鼠疫风险预测：利用指数平滑预测法对7项指标（$X_1\sim X_7$）进行趋势外推。指数平滑法的基本公式：$ES(t)=a*X(t-1)+(1-a)*ES(t-1),\cdots(*)$；$ES(t)$表示时间$t$的平滑值，本文$t$取值范围为1982—2012年；$X(t)$表示时间$t$的实际观测值；$a$为平滑常数，取值范围为[0,1]。

7项指标（$X_1\sim X_7$）的a值分别为0.81、0.45、0.44、0.94、0.73、0.97和0.76，采用上述a值计算得到的2012年7项指标（$X_1\sim X_7$）预测值分别为0.94、48%、1.85、64%、9.94、14.88%和0.28。由包含1～7个因子

表1 最优回归模型的系数及统计检验

因子个数	b_0	b_1	b_2	b_3	b_4	b_5	b_6	b_7	R^2	F值	P值
1	−0.098 6	−	−	0.152 2	−	−	−	−	0.195 9	6.820	0.014 3
2	−0.428 4	−	−	0.134 0	−	−	−	0.977 3	0.297 6	5.719	0.008 5
3	0.308 8	0.383 9	−0.021 7	0.241 8	−	−	−	−	0.389 5	5.530	0.004 5
4	1.032 1	0.555 5	−0.026 4	0.265 6	−0.011 6	−	−	−	0.472 7	5.603	0.002 3
5	0.947 9	0.515 4	−0.027 5	0.268 6	−0.011 4	−	0.010 7	−	0.485 8	4.536	0.004 7
6	0.962 8	0.558 9	−0.028 7	0.279 2	−0.010 1	−0.008 5	0.011 0	−	0.491 3	3.702	0.010 1
7	0.889 7	0.547 5	−0.027 0	0.269 6	−0.009 9	−0.008 6	0.007 8	0.167 3	0.492 5	3.050	0.021 1

注：b_0为常数项，$b_1\sim b_7$分别为$X_1\sim X_7$的回归系数估计

表2　选取不同数量指标的最优回归方程

因子个数	最优回归方程
1	$y=-0.098\,6+0.152\,2X_3$
2	$y=-0.428\,4+0.134\,0X_3+0.977\,3X_7$
3	$y=0.308\,8+0.383\,9X_1-0.021\,7X_2+0.241\,8X_3$
4	$y=1.032\,1+0.555\,5X_1-0.026\,4X_2+0.265\,6X_3-0.011\,6X_4$
5	$y=0.947\,9+0.515\,4X_1-0.027\,5X_2+0.268\,6X_3-0.011\,4X_4+0.010\,7X_6$
6	$y=0.962\,8+0.558\,9X_1-0.028\,7X_2+0.279\,2X_3-0.010\,1X_4-0.008\,5X_5+0.011\,0X_6$
7	$y=0.889\,7+0.547\,5X_1-0.027\,0X_2+0.269\,6X_3-0.009\,9X_4-0.008\,6X_5+0.007\,8X_6+0.167\,3X_7$

表3　最优回归方程不同因子数量的y值

年份	流行状况	1	2	3	4	5	6	7
1982	0	0.070 2	0.277 4	0.310 1	0.443 5	0.423 3	0.436 1	0.462 0
1983	0	0.114 4	0.384 7	-0.139 2	-0.048 0	-0.011 2	-0.012 6	0.033 2
1984	0	0.321 3	0.137 0	0.262 0	0.135 3	0.113 1	0.157 3	0.138 0
1985	1	0.220 9	0.185 3	0.226 3	0.449 5	0.425 0	0.436 6	0.432 3
1986	1	0.155 5	0.264 5	0.361 5	0.430 5	0.450 3	0.432 8	0.432 7
1987	1	0.196 5	0.300 7	0.624 2	0.667 3	0.689 5	0.713 0	0.700 1
1988	1	0.265 0	0.458 8	0.473 5	0.470 5	0.490 8	0.507 9	0.524 8
1989	1	0.252 8	0.252 6	0.528 4	0.608 9	0.527 8	0.539 0	0.547 7
1990	1	0.433 9	0.343 7	0.269 5	0.383 6	0.390 2	0.371 0	0.358 9
1991	0	0.227 0	0.161 4	0.179 5	0.029 2	0.014 8	0.024 6	0.024 1
1992	0	0.222 4	0.118 3	0.416 2	0.332 9	0.283 9	0.281 4	0.270 1
1993	0	0.430 9	0.331 2	0.515 9	0.412 3	0.376 7	0.432 8	0.425 2
1994	1	0.700 2	0.851 9	0.843 7	0.903 3	0.927 2	0.900 5	0.910 1
1995	0	0.433 9	0.431 6	0.196 2	0.375 8	0.360 0	0.351 1	0.363 5
1996	1	0.719 9	1.142 9	0.832 3	0.790 1	0.983 7	1.014 3	1.024 5
1997	0	0.520 6	0.605 8	0.653 2	0.647 7	0.636 9	0.642 2	0.653 5
1998	1	0.989 3	0.774 2	0.994 7	0.974 4	0.876 2	0.880 1	0.871 9
1999	0	0.373 0	0.309 6	0.253 1	0.420 4	0.420 4	0.323 7	0.315 5
2000	0	0.319 8	0.184 5	0.078 4	-0.106 7	-0.160 5	-0.218 7	-0.208 5
2001	0	0.173 7	0.085 2	-0.03 00	0.162 6	0.163 0	0.161 0	0.151 4
2002	0	0.062 6	0.221 9	0.361 4	0.298 2	0.346 1	0.324 9	0.322 3
2003	0	0.242 2	0.253 0	0.342 2	0.353 2	0.367 5	0.319 2	0.311 2
2004	0	0.164 6	0.155 3	0.316 0	-0.031 2	-0.016 7	-0.095 1	-0.101 3
2005	0	0.240 7	0.251 6	0.007 4	-0.109 6	-0.151 1	-0.121 7	-0.092 1
2006	0	0.246 7	0.198 4	-0.047 9	-0.238 8	-0.234 8	-0.216 6	-0.205 6
2007	0	0.204 1	0.092 4	-0.173 0	-0.160 3	-0.055 6	-0.044 1	-0.073 6
2008	0	0.150 9	0.133 5	0.151 3	0.145 0	0.148 2	0.166 6	0.162 6
2009	0	0.201 1	0.040 9	0.108 4	0.234 5	0.232 0	0.246 7	0.222 3
2010	0	0.152 4	-0.011 8	0.076 8	0.181 0	0.139 5	0.167 9	0.153 4
2011	0	0.193 5	0.063 5	0.007 7	-0.155 2	-0.156 4	-0.122 1	-0.130 2
P值[a]		0.066 8	0.005 0	0.001 3	<0.000 1	<0.000 1	<0.000 1	<0.000 1

注：[a] 采用Mann-Whitney u检验

表4　风险分级拟合率

因子个数	流行			不流行			高风险流行		
	预测值	实际值	拟合率(%)	预测值	实际值	拟合率(%)	预测值	实际值	拟合率(%)
1	3	3	100	17	22	77	1	5	20
2	3	3	100	18	22	82	2	5	40
3	3	3	100	16	18	89	4	9	44
4	4	4	100	15	15	100	5	11	45
5	4	4	100	14	14	100	5	12	42
6	4	4	100	17	17	100	5	9	56
7	4	4	100	17	17	100	5	9	56

的最优回归模型(1)～(7)计算得到的2012年y值分别为$y_1=0.183\,2$，$y_2=0.089\,0$，$y_3=0.084\,6$，$y_4=0.046\,3$，$y_5=0.049\,7$，$y_6=0.071\,2$，$y_7=0.060\,8$。按本研究提出的鼠疫风险分级标准，7个回归模型均预报未流行，预测2012年不会发生动物鼠疫流行。

讨　论

1. 鼠疫流行标准：本研究仅以检出鼠疫菌作为流行判定标准，而未将血清学阳性作为判定标准，如血凝抗体阳性取值为$y=0.5$时，在含有1～7个因子的最优回归模型中采用Mann-Whitney u检验（表5），当鼠疫不流行和血清学阳性时y值的差异无统计学意义，故不能通过风险分级方式将两者区分，本研究在风险分级中只选取检出鼠疫菌作为风险分级的标准。

2. 建立动物鼠疫流行风险分级回归方程：回归分析和时间序列预测是近年在预测、预警中常用的方法。最优回归子集方法可从一切可能有相同自变量个数的回归方程中选出相关系数最高者，其回归效果优于其他方法。指数平滑法强调时间序列的发展主要受近期数据的影响，任何一年的预测值都是本次实际观察值与前一年指数平滑值的加权平均，只需一个最新观察值、最新预测值和本年度的权重值就可以进行趋势预测。因此与全期平均法和移动平均法相比，

表5 检出鼠疫菌和血凝抗体阳性作为风险分级标准的Mann-Whitney u检验

年份	流行状况	1	2	3	4	5	6	7
1982	0	0.376 1	0.429 5	0.541 2	0.598 3	0.577 5	0.573 6	0.566 2
1983	0	0.409 5	0.427 3	0.252 8	0.291 9	0.295 4	0.302 9	0.289 9
1984	0.5	0.566 1	0.662 9	0.551 8	0.497 5	0.427 4	0.423 4	0.428 9
1985	1	0.490 1	0.477 8	0.490 6	0.586 2	0.567 0	0.562 2	0.563 4
1986	1	0.440 6	0.549 3	0.597 6	0.627 1	0.655 4	0.659 2	0.659 2
1987	1	0.471 7	0.716 4	0.802 1	0.820 6	0.784 5	0.789 3	0.793 0
1988	1	0.523 5	0.503 9	0.651 1	0.649 9	0.623 8	0.628 1	0.623 4
1989	1	0.514 3	0.451 8	0.674 2	0.708 7	0.688 1	0.671 7	0.669 2
1990	1	0.651 4	0.683 6	0.554 8	0.603 7	0.634 1	0.635 2	0.638 7
1991	0.5	0.494 7	0.424 0	0.447 5	0.383 1	0.367 3	0.364 5	0.364 6
1992	0.5	0.491 3	0.530 5	0.623 7	0.588 0	0.590 1	0.580 1	0.583 3
1993	0	0.649 1	0.578 4	0.686 0	0.641 7	0.552 4	0.545 8	0.547 9
1994	1	0.853 0	0.922 7	0.960 9	0.986 4	1.029 2	1.033 8	1.031 1
1995	0.5	0.651 4	0.578 1	0.482 8	0.559 8	0.573 2	0.569 8	0.566 3
1996	1	0.867 9	0.877 8	0.941 6	0.923 6	0.882 8	0.922 7	0.919 8
1997	0.5	0.717 0	0.721 1	0.802 2	0.799 8	0.791 0	0.788 9	0.785 7
1998	1	1.071 8	1.194 9	1.105 0	1.096 2	1.086 5	1.066 4	1.068 8
1999	0.5	0.605 3	0.459 7	0.494 1	0.565 7	0.717 3	0.716 3	0.718 6
2000	0.5	0.565 0	0.407 6	0.373 8	0.294 5	0.383 9	0.372 2	0.369 3
2001	0.5	0.454 4	0.462 0	0.326 9	0.409 4	0.412 6	0.412 7	0.415 4
2002	0.5	0.370 3	0.678 9	0.634 3	0.607 2	0.642 2	0.651 8	0.652 5
2003	0.5	0.506 2	0.520 8	0.573 2	0.577 9	0.654 1	0.656 5	0.658 8
2004	0.5	0.447 5	0.542 2	0.566 4	0.417 7	0.541 1	0.543 2	0.544 9
2005	0.5	0.505 1	0.417 7	0.335 9	0.285 8	0.238 2	0.230 1	0.221 7
2006	0.5	0.509 7	0.366 3	0.288 1	0.206 3	0.178 0	0.179 0	0.175 9
2007	0	0.477 4	0.174 4	0.165 0	0.170 4	0.156 1	0.177 6	0.186 0
2008	0.5	0.437 1	0.486 2	0.449 2	0.446 5	0.417 7	0.418 5	0.419 7
2009	0.5	0.475 1	0.376 3	0.392 5	0.446 5	0.423 5	0.423 1	0.430 0
2010	0	0.438 3	0.464 2	0.396 5	0.441 2	0.395 1	0.387 0	0.391 1
2011	0	0.469 4	0.413 5	0.338 0	0.268 2	0.214 4	0.214 5	0.216 8
P值(血凝抗体阳性*vs.* 动物检菌阳性)		0.199 7	0.031 8	0.003 5	0.000 4	0.004 2	0.003 5	0.003 5
P值(血凝抗体阳性*vs.* 未流行)		0.149 8	0.330 5	0.293 3	0.508 1	0.149 8	0.149 8	0.149 8
P值(动物检菌阳性*vs.* 未流行)		0.025 6	0.007 6	0.012 0	0.004 8	0.000 8	0.000 8	0.000 8

注：因变量取值动物检菌为$y=1$，血凝抗体阳性为$y=0.5$，未流行为$y=0$

既不需要存储全部历史数据，也无需存储一组数据，从而可以大大减少数据存储的问题。

动物鼠疫的发生与许多因素相关，分析时应筛选出主要的影响因素。米景川等[4]根据1982—1993年内蒙古北部荒漠草原区的天文、气象和监测数据建立了多元回归和最优回归模型预测动物鼠疫流行强度，应用总体模型对3个阶段36年次的流行强度拟合，总体拟合率为86.11%，分段拟合率均达到100%。随着自然环境的改变，达乌尔黄鼠密度也在变化，但连续几年的数据难有大的变化，因此在数据分析中利用近年的数据可以降低历史数据对预测结果的影响，本研究利用历史观测数据，并采用指数平滑方法计算历史数据预测值和观测值方差最小的权重值，由此进行趋势预测[5]。

参 考 文 献

[1] Zhang CH, Lyu JS, Pu QJ. The present situation and control countermeasure of Spermophilus Dauricus Plague Natural Foci [J]. Chin J Contl Endem Dis, 2004, 19(6): 345-348. (in Chinese)
张春华，吕景生，浦清江. 达乌尔黄鼠疫源地的现状及防治对策[J]. 中国地方病防治杂志，2004，19(6)：345-348.

[2] Fang XY. China plague natural foci [M]. Beijing: People's Medical Publishing House, 1990: 116-135. (in Chinese)
方喜业. 中国鼠疫自然疫源地[M]. 北京：人民卫生出版社，1990：116-135.

[3] Zhu JQ, Wang WS, Zhang HY, et al. Study on the criteria for determinatong plague natural foci and plague epizootics [J]. Endem Dise Bull, 1996, 11(1): 78-80. (in Chinese)
朱锦沁，汪闻绍，张鸿猷，等. 鼠疫自然疫源地及动物鼠疫流行判定标准的研制[J]. 地方病通报，1996，11(1)：78-80.

[4] Mi JC, Li ZL, Lyu WD, et al. Northern desert grassland of Inner Mongolia gerbil plague epidemic intensity forecast model [J]. Chin J Contl Endem Dis, 1997, 12(2): 109-112. (in Chinese)
米景川，李仲来，吕卫东，等. 内蒙古北部荒漠草原沙鼠鼠疫流行强度的预报模型[J]. 中国地方病防治杂志，1997，12(2)：109-112.

[5] Zeng ZY, Yang YM, Song ZM, et al. A comparative study of biomass dynamics of the species populations of the desert community in the Chihuahuan desert of North America [J]. Zool Res, 1994, 15(2): 32-41. (in Chinese)
曾宗永，杨跃敏，宋志明，等. 北美Chihuahuan荒漠啮齿动物群落各物种种群生物量动态的比较研究[J]. 动物学研究，1994，15(2)：32-41.

（收稿日期：2013-08-09）
（本文编辑：张林东）