基于多元统计分析的

Montesinho 自然公园森林火险预警方法初探

Xiaozhu Zhang

Hunan University

Jan. 8, 2019

Contents

1 问题的简述

1.1 背景和数据

森林大火是一种突发性强、破坏性大、处置救助较为困难的自然灾害，对生态系统、人类的生命财产安全等都构成了较大的威胁。如果能够根据森林火灾发生前的种种气象特征，对火灾发生与否、火灾严重程度做出推断，并且建立相应的等级预报方法，那么森林火灾的预防和控制将会更加有效。

本文的数据集来自网站UCI Machine Learning Repository, 访问网址为`https://archive.ics.uci.edu/ml/datasets/Forest+Fires`. 数据集记录的是位于葡萄牙东北部山地Trás-os-Montes的Montesinho自然公园的森林火灾数据。这个国家公园拥有丰富的动植物群落，年均温度在8—12℃，却时常受到森林火灾的威胁，数据记录十分完整，没有缺失值。

数据集中一共有517个样本，时间跨度在2000年1月至2003年12月。13个变量分别是：

X	观测点的横坐标，1 to 9;
Y	观测点的纵坐标，2 to 9;
$month$	月份，“jan” to “dec”;
day	日期，“mon” to “sun”;
$FFMC$	FFMC指数，18.70 to 96.20;
DMC	DMC指数，1.1 to 291.3;
DC	DC指数，7.9 to 860.6;
ISI	ISI指数，0.00 to 56.10;
$temp$	摄氏度(℃)，2.2 to 33.3;
RH	相对湿度(%)，15.0 to 100.0;
$wind$	风速(km/h)，0.4 to 9.4;
$rain$	降雨量(mm/m^2)，0.0 to 6.4;
$area$	过火面积(ha)，0.00 to 1090.84.

其中，FFMC指数、DMC指数、DC指数和ISI指数是加拿大森林火险气候指数(FWI)系统内部的指标，可以综合反映各项影响森林火灾发生与蔓延的气象条件。

本文的分析将基于R语言进行。

```
> data=read.csv("forestfires.csv")
> head(data,5)

  X Y month day FFMC  DMC    DC ISI temp RH wind rain area
1 7 5   mar fri 86.2 26.2  94.3 5.1  8.2 51  6.7  0.0    0
2 7 4   oct tue 90.6 35.4 669.1 6.7 18.0 33  0.9  0.0    0
3 7 4   oct sat 90.6 43.7 686.9 6.7 14.6 33  1.3  0.0    0
4 8 6   mar fri 91.7 33.3  77.5 9.0  8.3 97  4.0  0.2    0
5 8 6   mar sun 89.3 51.3 102.2 9.6 11.4 99  1.8  0.0    0

> attach(data)
```

1.2 问题的拆分

本文的目的，是尝试建立基于“forestfires”数据集的森林火险预警方法。为了更有条理地解决问题，拟从以下几个步骤进行分析：

（1）从观测点坐标的分布，说明数据来源的合理性；

（2）对过火面积$area$进行聚类，将其分为不同的等级；

（3）利用典型相关分析，探究FFMC指数、DMC指数、DC指数和ISI指数与气象条件之间的关系；

（4）利用对应分析，探究月份与火灾等级之间的关系；

（5）利用对数线性模型，探究工作日/休息日与火灾严重/不严重之间的关系；

（6）利用判别分析，初步建立森林火险等级预报模型；

（7）利用Logistic模型，对（6）中建立的模型进行改进。

2 模型的建立与分析

2.1 观测点坐标分布

首先，以X为横坐标，Y为纵坐标，绘制出观测点的位置分布图。

```
> library(ggplot2)
> position=data.frame(X,Y)
> ggplot(position,aes(x = X, y = Y))+
+   geom_point(size=2)+
+   labs(x = "x-axis spatial coordinate", y = "y-axis spatial coordinate")+
+   scale_x_continuous(breaks=seq(1,9,1))+
+   scale_y_continuous(breaks=seq(2,9,1))
```

散点图如Figure 1所示。

不难看出，各个观测点的位置分布较为均匀，因此本数据集能够反映自然公园内各个位置的特征，利用此数据进行预测是比较适合的。

2.2 过火面积$area$的聚类分析

下面，做出过火面积$area$的直方图。

```
> ggplot(data,aes(x=area))+geom_histogram()
```

观察Figure 2，会发现变量$area$极度右偏。因此，要对这样的数据进行聚类，十分自然地，我们会考虑“极端值”的问题，并对其加以利用。

受到描述统计学中“五数概括法”(five-number summary)的启发，此处，不妨通过判断“上极端值”的方法对过火面积$area$进行等级划分。相较于一般的聚类方法，这种方法或许能够更加充分地利用数据分布的特点，得到合理的聚类结果。

具体算法如下：

（1）将$area$数据加载进入程序；

（2）将程序内数据从小到大进行排序；

（3）以最底层数据（最小值）为起点，每次向上选入一个数据；

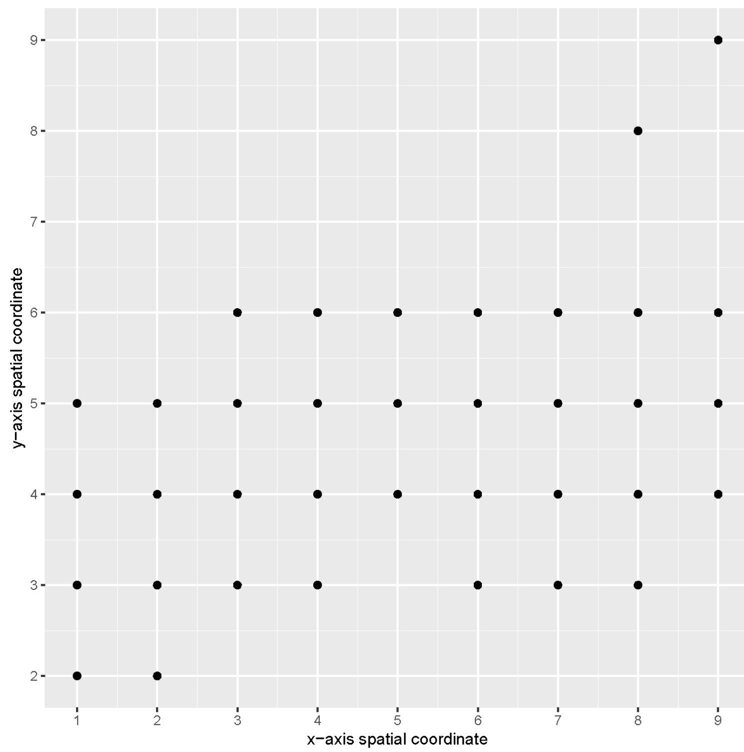

Figure 1: 观测点坐标

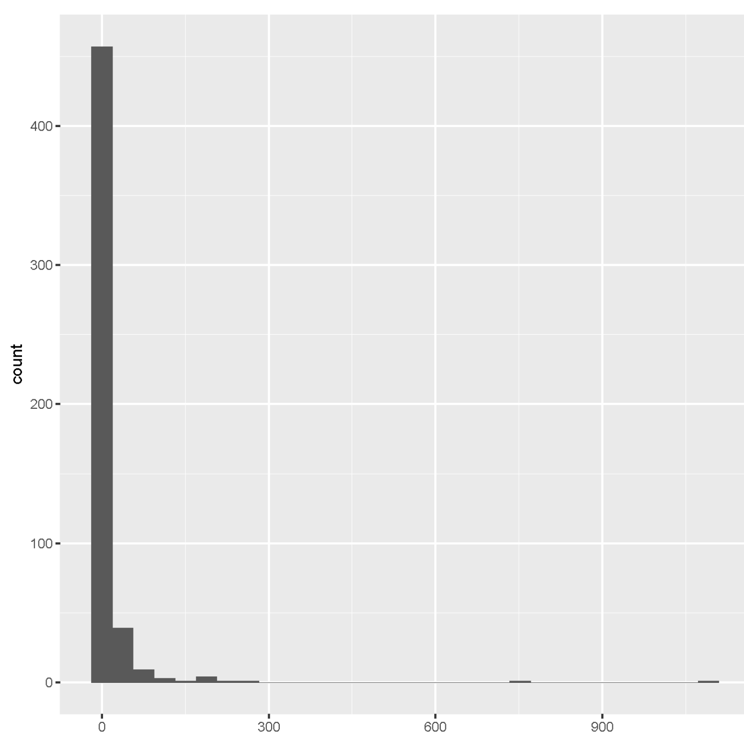

Figure 2: $area$直方图

（4）计算被选入数据的25%分位点Q_1，75%分位点Q_2，以及最大值max；

（5）若$(max - Q_3) \geq 1.5 \cdot IQR$，则认为存在上极端值，将已选入的数据归为一类，抛出程序，并继续重复（2）；

（6）若$(max - Q_3) < 1.5 \cdot IQR$，则认为不存在上极端值，并继续重复（3）.

```
> A=sort(area)
> k=1.5 #设定临界值
> cluster=c()
> type=1
> start=1
> for(i in 1:length(A)){
+   q=quantile(A[start:i])
+   if(q[5]-q[4]<=k*(q[4]-q[2])){
+     cluster=c(cluster,type)
+   }else{
+     type=type+1
+     cluster=c(cluster,type)
+     start=i+1
+   }
+ }
> degree=c()
> for(i in 1:length(area)){
+   for(j in 1:length(area)){
+     if(area[i]==A[j]){
+       degree[i]=cluster[j]
+     }
+   }
+ }
```

经过计算，最终过火面积$area$被分为了6个等级。

Table 1: 过火面积$area$的聚类结果

等级	1	2	3	4	5	6
过火面积	0	0.09 to 15.64	16.0 to 71.3	82.8 to 105.4	154.9 to 212.9	278.5 to 1090.8
样本数量	247	204	51	6	6	3

不难发现，所有过火面积$area$为0的样本，被单独聚为一类，记作“等级1”。“等级2”和“等级3”的过火面积较小，可以被看作轻微火灾；而“等级4”、“等级5”和“等级6”的过火面积较大，可以被看做重度火灾。其中“等级6”的过火面积跨度较大，最大值破千，是“超级火灾”，需要格外引起防控人员的注意。

由于过火面积$area$的分布距离正态分布相去甚远，并不适合进行多元分析，因此将其转化为各个等级这样的定性数据。此外，随着过火面积的增大，等级的提高，每一等级的样本数量呈骤减趋势，这对后续的多元分析也是尤为不利的；因此，在后续的分析中，会将“等级4”、“等级5”和“等级6”合并，如此合并后的新类，其内部的样本数量才可与其它类别勉强抗衡。

采用这种方式进行聚类，最大的优点就是每一别类内部都没有极端值，对严重偏态的数据进行了恰当的处理。通过箱线图Figure 3，可以更加形象地了解6个等级的特点。

```
> data<-cbind(data,degree)
> #boxplot after cluster
> ggplot(data,aes(x=as.factor(degree),y=area))+
+   geom_boxplot()+
+   labs(x = "degree", y = "area")
```

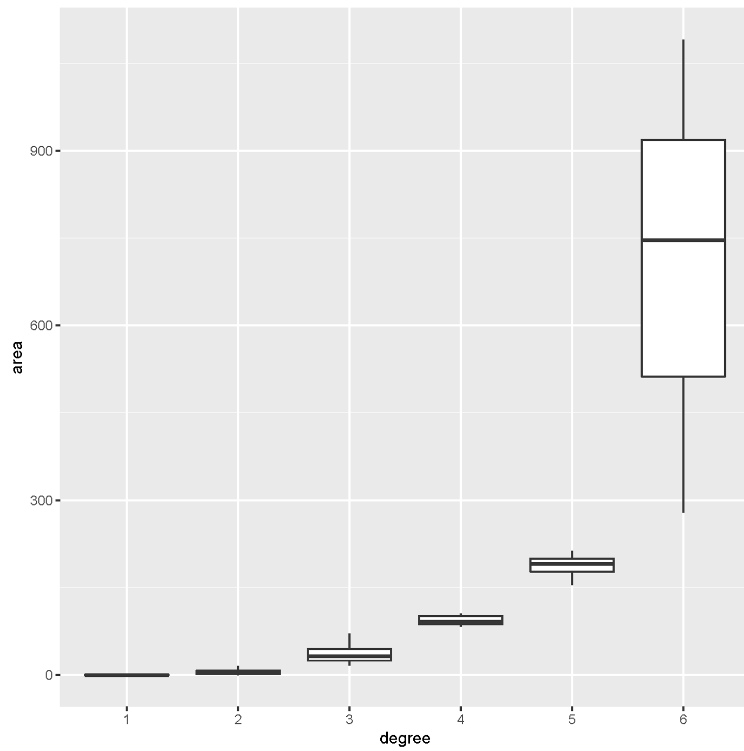

Figure 3: 6个等级的箱线图

2.3 FWI系统指数与气象指标的典型相关分析

通过查阅资料，了解到FFMC指数、DMC指数、DC指数和ISI指数是由气象指标$temp$、RH、$wind$和$rain$经过复杂的计算得到的，是气象条件的一种综合反映。那么，这4个FWI指标和气象条件指标是如何联系的，联系程度有多高？此处采用典型相关分析进行说明。

```
> data.X=data[,5:8]
> data.Y=data[,9:12]
> can=cancor(data.X,data.Y)
```

经过计算，4对典型变量的典型相关系数分别为

```
> can$cor
```

```
[1] 0.66790976 0.37712731 0.24674008 0.03043379
```

下面，对典型相关系数进行显著性检验，选出显著的典型变量。

```
> #序贯检验
> cor.test<-function(X,Y,alpha){
+   n=nrow(X)
+   p=ncol(X)
+   q=ncol(Y)
+   can<-cancor(X,Y)
+   j=min(p,q)
+   count=0
+   QSTA=c()
+   PVAL=c()
+   for(i in 1:j){
+     Lambda=prod(1-can$cor[i:j]^2)
+     Q=-(n-i-1/2*(p+q+1))*log(Lambda)
+     pvalue=pchisq(Q,(p-i+1)*(q-i+1),lower.tail = F)
+     if(pvalue<alpha){
+       count=count+1
+     }
+     QSTA=c(QSTA,Q)
+     PVAL=c(PVAL,pvalue)
+   }
+   return(list("count"=count,"Q-statistics"=QSTA,"pvalue"=PVAL))
+ }

> #test
> cor.test(data.X,data.Y,alpha=0.05)

$count
[1] 3

$`Q-statistics`
[1] 413.2562840 110.8565194  32.4752072   0.4711988

$pvalue
[1] 6.044023e-78 9.840421e-20 1.529591e-06 4.924362e-01
```

根据计算结果，一共有3对典型变量是显著的。因此，在后续的分析中，只考虑前3对典型变量。

典型变量U和V的系数分别为：

```
> can$xcoef[,1:3]

              [,1]          [,2]          [,3]
FFMC  2.282507e-04  8.004781e-03  0.0021946785
DMC  -3.587303e-04 -5.423785e-04 -0.0002499374
DC   -7.059919e-05  4.112021e-05  0.0001553032
ISI  -3.518497e-03  2.419986e-04 -0.0086344042

> can$ycoef[,1:3]

             [,1]          [,2]          [,3]
temp -0.009107003 -0.0002420258 -0.0012253194
RH   -0.001634249 -0.0027599127 -0.0002974532
wind -0.002184744  0.0020418450 -0.0248029932
rain  0.005212364  0.0178705316 -0.0120657261
```

也就是，

$$\begin{cases} U_1 = 2.28*10^{-4}\cdot FFMC - 3.59*10^{-4}\cdot DMC - 7.06*10^{-5}\cdot DC - 3.52*10^{-3}\cdot ISI \\ V_1 = -0.009\cdot temp - 0.002\cdot RH - 0.002\cdot wind + 0.005\cdot rain \end{cases}$$

$$\begin{cases} U_2 = 8.00*10^{-3}\cdot FFMC - 5.42*10^{-4}\cdot DMC - 4.11*10^{-5}\cdot DC - 2.42*10^{-4}\cdot ISI \\ V_2 = -0.0002\cdot temp - 0.0028\cdot RH - 0.0020\cdot wind + 0.0179\cdot rain \end{cases}$$

$$\begin{cases} U_3 = 0.0022\cdot FFMC - 0.00025\cdot DMC + 0.00016\cdot DC - 0.0086\cdot ISI \\ V_3 = -0.0012\cdot temp - 0.0003\cdot RH - 0.0248\cdot wind - 0.0121\cdot rain \end{cases}$$

然而，在分析典型变量和原始变量之间的关系时，典型变量的系数并不是一个好的工具，其意义更多地在于帮助计算典型变量的值。

为了更加合理地说明变量之间的关系，此处采用“典型载荷”，即典型变量和原始变量之间的相关系数。

```
> U=as.matrix(data.X)%*%as.matrix(can$xcoef[,1:3])
> colnames(U)=paste("U",1:3,sep="")
> V=as.matrix(data.Y)%*%as.matrix(can$ycoef[,1:3])
> colnames(V)=paste("V",1:3,sep="")
> #典型载荷
> cor(data.X,U)

             U1          U2          U3
FFMC -0.4963521  0.79173678 -0.05027227
DMC  -0.8935364 -0.23931278  0.06581029
DC   -0.8279094  0.03089804  0.51309838
ISI  -0.5996047  0.37118565 -0.65833805

> cor(data.Y,V)

               V1          V2          V3
temp -0.85912662  0.49706890  0.1201506
RH    0.02508469 -0.98840539 -0.1031743
wind  0.14399804  0.02668728 -0.9853490
rain -0.11429061  0.02095750 -0.1650420
```

根据计算结果，第一典型变量U_1与DMC、DC密切相关，V_1与$temp$密切相关；第二典型变量U_2与$FFMC$密切相关，V_2与RH密切相关；而第三典型变量U_3与DC、ISI密切相关，V_3则与$wind$密切相关。

为了更加形象地展示变量之间的关系，将每一组的原始变量与其典型变量对应，根据成对典型变量之间的相关关系，可以做出关系图谱。

Figure 4展示了“FWI系统指数”通过其组内典型变量U_i、对应的典型变量V_i的传递，与“气象指标”之间建立的关系。其中，“+”代表正相关，“-”代表负相关；而上方有“（主）”这一符号的变量，是与前一箭头尾端的变量相关系数（的绝对值）最

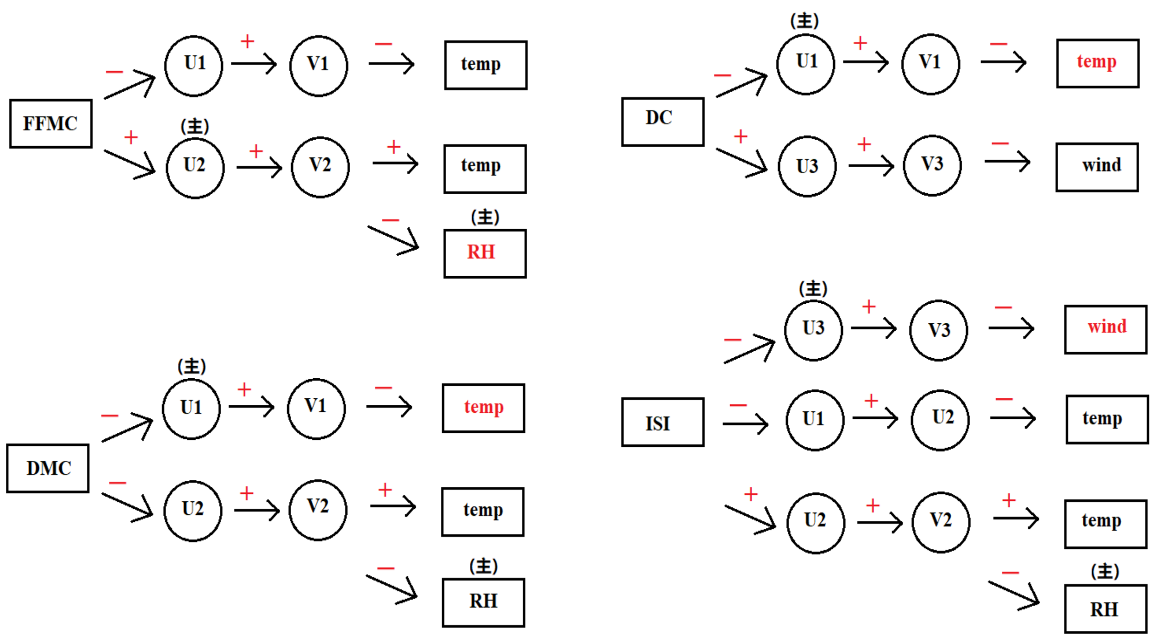

Figure 4: 变量之间的关系

大的变量。通过寻找主要路径，在图谱的末端用红色标出了与图谱首端的"FWI系统指数"变量相关性最强的"气象指标"变量。

具体地，$FFMC$主要与相对湿度RH相关，RH越高，$FFMC$越小；同时，$FFMC$也与温度$temp$存在一定的正相关。

DMC主要与温度$temp$相关，$temp$越高，DMC越大；并且DMC与相对湿度RH也存在一定的正相关。

DC主要与温度$temp$相关，$temp$越高，DC越大；此外，DC与风速$wind$存在一定的负相关。

影响ISI的因素最为复杂。它主要与风速$wind$相关，$wind$越高，ISI也越高；同时，$temp$和RH也在一定程度上影响ISI，$temp$越高，ISI越大；而RH越高，ISI反而越小。

结合上述分析，通过查阅文献资料 [1]，得知，$FFMC$反映的是"细小可燃物湿度"，DMC反映的是"粗腐殖湿度"，DC反映的是"干旱"，ISI反映的是"初始蔓延"。所有的"FWI系统指数"都遵循"越大代表火险越高"的规律，而这与典型相关分析的结果是完全相合的。

事实上，"FWI系统指数"比温度、湿度、风速、降雨等"气象指标"具有更重要的意义，对后续分析也更加友好。这是因为，"气象指标"的量纲不一致，并且某些指标提供的信息过少。例如，数据集中的"降雨"$rain$变量，仅有8个样本不为0，进而导致协方差矩阵是退化的；而FWI系统指数信息充足、在取值范围上相对一致，更加适合用作建模。

基于上述考虑，在后续的分析中，将主要采用"FWI系统指数"的4个变量，对过火面积$area$的等级$degree$进行预测。

2.4 月份与火灾等级的对应分析

众所周知，森林火灾较易发生在夏季，并且不同月份的灾情存在着显著差异。

```
> t=table(month,degree)
> t

     degree
month  1  2  3  4  5  6
  apr  5  3  1  0  0  0
  aug 85 79 15  1  3  1
  dec  0  6  3  0  0  0
  feb 10  8  2  0  0  0
  jan  2  0  0  0  0  0
  jul 14 15  1  1  0  1
  jun  9  7  1  0  0  0
  mar 35 14  5  0  0  0
  may  1  0  1  0  0  0
  nov  1  0  0  0  0  0
  oct 10  3  2  0  0  0
  sep 75 69 20  4  3  1
```

根据月份和火灾等级的列联表，火灾频繁出现在7月至9月，且3月份也频繁出现“轻微火灾”。其中，“重度火灾”和“超级火灾”只发生在7月至9月，这一时期正是葡萄牙的炎炎夏季。

然而，由于各个月份的样本数不相等，在某些月份样本较多，而另一些月份样本较少，因此仅仅对列联表进行描述性分析极易产生错误。下面，从行剖面和列剖面的角度，将各个状态的边缘概率纳入考虑范围，得出更加合理的结论。

```
> #行剖面
> tmonth=c()
> tD=c()
> freq=c()
> pt=prop.table(t,1)
> for(i in 1:nrow(t)){
+   for(j in 1:ncol(t)){
+     tmonth=c(tmonth,rownames(t)[i])
+     tD=c(tD,colnames(t)[j])
+     freq=c(freq,t[i,j])
+   }
+ }
> ggplot(data.frame(month=tmonth,degree=tD,percentage=freq),
+        aes(month,percentage,fill=degree))+
+   geom_bar(stat="identity",position="fill")
```

Figure 5是行剖面的百分比簇状条形图。从行剖面（各月发生火灾情况）的角度观察，无论是哪一个月份，“等级1”和“等级2”的火灾比例都是最大的；即使是“重度火灾”高发的7月至9月，“重度火灾”也仅仅占到很小的比例。当然，这并不是可以忽视夏季火险的理由，毕竟“重度火灾”的过火面积破百甚至破千，一旦发生就会损失惨重。

此外，1月和11月完全没有发生火灾。12月的“轻微火灾”比例不小，这或许与12月节日（如圣诞节）的人为活动有关。

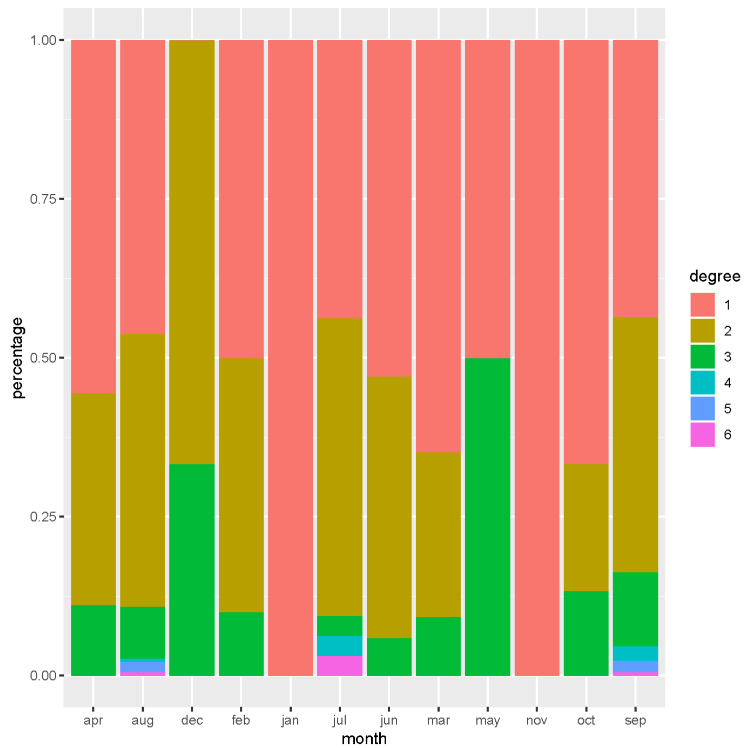

Figure 5: 行剖面（各月发生火灾情况）

```
> #列剖面
> ggplot(data.frame(month=tmonth,degree=tD,percentage=freq),
+        aes(degree,percentage,fill=month))+
+   geom_bar(stat="identity",position="fill")
```

Figure 6是列剖面的百分比簇状条形图。从列剖面（各级火灾的月份分布情况）的角度观察，对于任何级别的火灾，8月和9月所占的比例都是最大的；随着级别逐渐提高，每一个条图内部的色块（月份）越来越少——“等级4”、“等级5”和“等级6”的火灾，几乎全部发生在7月至9月。这说明，7月至9月的火灾特点一方面在于火灾级别之高，另一方面在于小规模火灾发生的频繁。

最后，将行剖面和列剖面的信息进行综合分析，即进行对应分析。

```
> #对应分析
> library(ca)
> ca.month=ca(table(month,degree))
> plot(ca.month)
```

在对应分析图Figure 7中，距离“等级4”、“等级5”和“等级6”最近的月份，正如我们所料，是7月、8月和9月；此外，8月和9月距离“等级1”、“等级2”和“等级3”也较近，说明除了“重度火灾”，在8月和9月也频繁地出现小规模的“轻度火灾”。

对于其他的月份，1月和11月（在图中重叠）距离“等级1”最近，距离所有其他过火面积大于0的等级最远，说明1月和11月是最不易发生发生森林火灾的月份。其

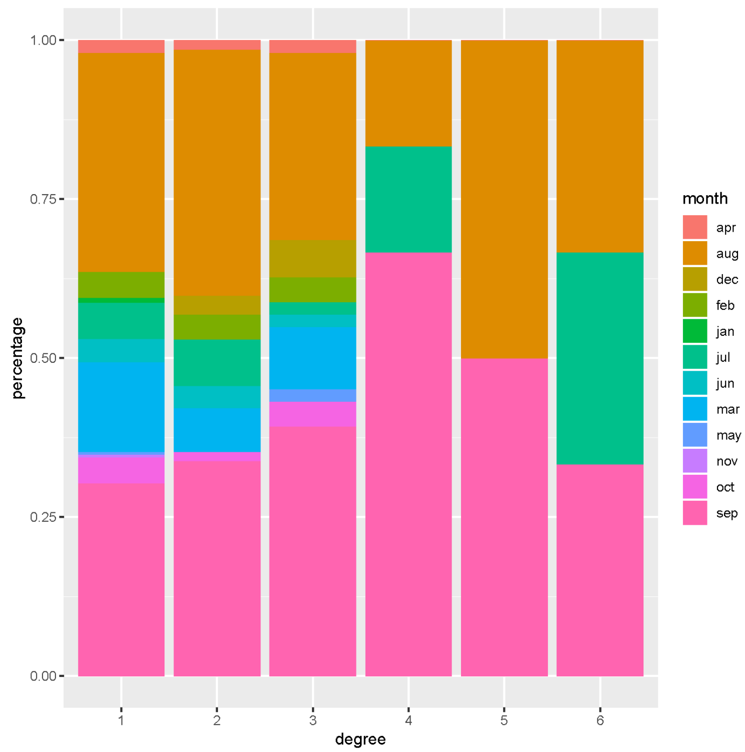

Figure 6: 列剖面（各级火灾的月份分布情况）

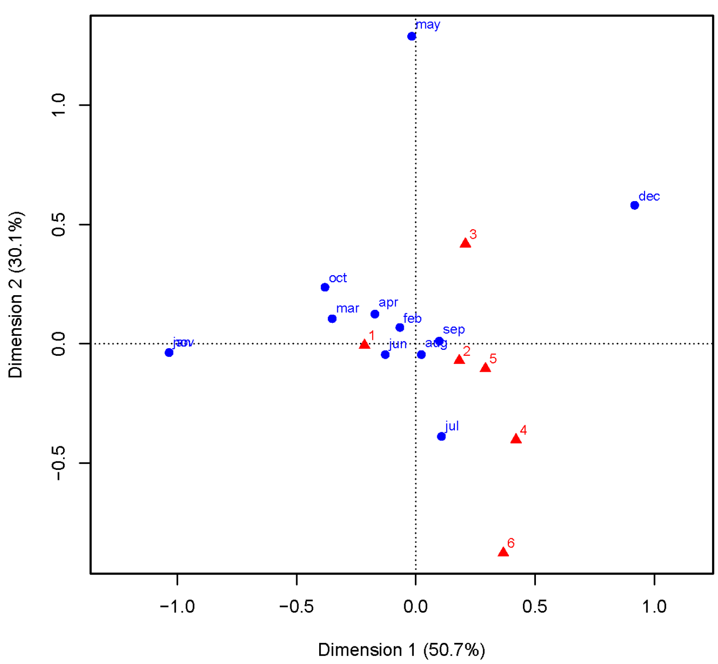

Figure 7: “月份”和“火灾等级”的对应分析图

他的春秋两季月份，密集地环绕在“等级1”和“等级2”周围，说明它们并非森林火灾的高发时段。

最后，需要着重关注12月这个离群点。12月距离“等级2”和“等级3”最近，说明尽管处于严冬，12月仍频繁发生小规模火灾。与上述“行剖面”的分析一致，这种现象包含节日期间的人为因素。

综上所述，对于森林火灾的防控人员而言，7月、8月和9月是需要着重关注的月份，期间可能伴随着频繁的“轻微火灾”和过火面积极大的“重度火灾”、“超级火灾”，因此整个夏季都不得松懈。同时，12月也是需要重点防控的月份，严禁游人携带火种，或许可以有效降低火险。

2.5 工作日/休息日与火灾严重程度的对数线性模型

根据Cortez和Morais [2]，森林火灾的发生和工作日/休息日（weekday/weekend）也有着密切的关联。与12月节日期间的火灾类似，在休息日期间，国家公园的客流量增大，人为活动的增多往往会增大火灾发生的风险。基于这种考虑，首先做出列联表进行观察。

```
> table(degree,day)

      day
degree fri mon sat sun thu tue wed
     1  42  35  42  48  30  28  22
     2  33  29  30  34  26  26  26
     3  10   9   8  11   4   6   3
     4   0   0   0   1   0   3   2
     5   0   0   3   1   0   1   1
     6   0   1   1   0   1   0   0
```

不难发现，周五、周六和周日的火灾数量的确比工作日的数量高一些。但是，这种情况仅仅针对小规模、低级别火灾。对于“重度火灾”、“超级火灾”，则没有很明显的规律可循。这主要是因为，人为因素（游客）造成的火灾往往能够得到及时的控制，不会大范围蔓延；而大范围蔓延的火灾往往是因为找不到起火原点而失去控制。

下面，将“等级1”至“等级3”进行合并，称之为“不严重火灾”，将“等级4”至“等级6”合并，称之为“严重火灾”。再将日期按休息日/工作日进行合并。做出2×2列联表进一步观察。

```
> serious<-c()
> weekend<-c()
> for(i in 1:nrow(data)){
+   if(day[i]=="fri" || day[i]=="sat" || day[i]=="sun"){
+     weekend=c(weekend,1)
+   }else{
+     weekend=c(weekend,0)
+   }
+   if(degree[i]==4 || degree[i]==5 || degree[i]==6){
+     serious=c(serious,1)
+   }else{
+     serious=c(serious,0)
```

```
+   }
+ }
> day.test<-data.frame(weekend,serious)
> table(day.test)

       serious
weekend   0   1
      0 244   9
      1 258   6
```

与上述分析一致，在“serious”为0这一列，休息日的样本数超过了工作日的样本数，说明不严重的火灾更多地发生在周末；而在在“serious”为1这一列，工作日的样本数超过了休息日的样本数，说明严重的火灾更多地发生在工作日。

然而，这样的差别是否显著？下面，建立对数线性模型进行分析：

$$\ln n_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

其中μ表示平均效应，α_i表示工作日/休息日的效应，β_j表示火灾不严重/严重的效应，γ_{ij}表示变量间的交互作用。

```
> library(MASS)
> logmol=loglm(~weekend*serious,table(day.test))
> logmol$param

$`(Intercept)`
[1] 3.759778

$weekend
          0           1
 0.08741844 -0.08741844

$serious
        0         1
 1.765286 -1.765286

$weekend.serious
       serious
weekend          0          1
      0 -0.1153141  0.1153141
      1  0.1153141 -0.1153141
```

根据分析的结果，在所有的样本中，工作日的占比更大，不严重火灾的占比更大。分析它们的交互效应，与先前的猜想一致，“工作日⇔严重火灾”与“休息日⇔不严重火灾”这两组关系存在正交互效应。

但是，当对交互效应进行显著性检验时，

```
> summary(table(day.test))

Number of cases in table: 517
Number of factors: 2
Test for independence of all factors:
        Chisq = 0.7567, df = 1, p-value = 0.3844
```

Pearson卡方统计量却很小，说明两个变量是相互独立的。

综合上述的种种分析，我们可以这样认为：在统计意义上，“工作日/休息日”与“火灾严重/不严重”是相互独立的。但是在现实防控的过程中，由于各种人为因素，仍然不能忽视休息日期间的火灾风险，而这一点对于无需被防控人员着重关注的10月至次年6月而言，格外重要。

2.6 基于FWI系统指数的火险等级判别分析

如前文所述，FWI系统指数是对气象特征指标的一种综合，具有量纲一致、信息充足的特点，更加适合用作多元分析。本文的最终目的，是尝试建立森林火险等级的预报方法。因此，在这一节，将以火灾等级$degree$为被解释变量，以$FFMC$、DMC、DC和ISI为判别变量，进行判别分析。

此处采用的是“距离判别法”。[1]

Step 1. 等级的合并

如2.2节中所述，随着过火面积的增大，等级的提高，每一等级的样本数量呈骤减趋势。为了更好地进行判别分析，下面，将“等级4”、“等级5”和“等级6”合并，并计算每一个新等级的样本量。

```
> #将4、5、6进行合并(因为样本数量过少)
> r=4
> G1 <- subset(data,degree==1)
> G2 <- subset(data,degree==2)
> G3 <- subset(data,degree==3)
> G4 <- rbind(subset(data,degree==4),
+             subset(data,degree==5),
+             subset(data,degree==6))
> n1=nrow(G1)
> n2=nrow(G2)
> n3=nrow(G3)
> n4=nrow(G4)
> n=n1+n2+n3+n4
> data.frame(n1,n2,n3,n4,n)

   n1  n2 n3 n4   n
1 247 204 51 15 517
```

合并后，新得到的“等级4”内共有15个单位，完全满足$n > p$的要求。

Step 2. 方差阵的同质性检验

在利用距离判别法分析之前，需要确定4个子总体的协方差阵是否相同。因此，构造F统计量进行检验。

[1]此处，可行的判别分析方法还包括“线性判别法”和“二次判别法”。但是，经过尝试，这两种方法效果均不佳。其中，“线性判别法”的判别精度高于“距离判别法”，为48.55%，但是没有任何样本被判别为“等级3”或“等级4”；“二次判别法”的精度低于“距离判别法”，为41.97%，并且没有任何样本被判别为“等级3”。
代码详见附录。

```
> #判断协方差阵是否同质
> cov.test <- function(p,nr){ #输入c(检验变量个数,检验变量编号(向量))
+   #构造M统计量
+   L1=(n1-1)*cov(G1[,nr])
+   L2=(n2-1)*cov(G2[,nr])
+   L3=(n3-1)*cov(G3[,nr])
+   L4=(n4-1)*cov(G4[,nr])
+   L=L1+L2+L3+L4
+   A=(n-r)*log(det(L/(n-r)))
+   B1=(n1-1)*log(det(L1/(n1-1)))
+   B2=(n2-1)*log(det(L2/(n2-1)))
+   B3=(n3-1)*log(det(L3/(n3-1)))
+   B4=(n4-1)*log(det(L4/(n4-1)))
+   M=A-B1-B2-B3-B4
+   #寻找分布
+   if(n1==n2 && n2==n3 && n3==n4){
+     d1=(2*p^2+3*p-1)*(r-1)/(6*(p+1)*r*(n-1))
+     d2=(p-1)*(p+2)*(r^2+r+1)/(6*r^2*(n-1)^2)
+   }
+   if(n1!=n2 || n1!=n3 || n2!=n3 || n3!=n4){
+     d1=(2*p^2+3*p-1)/(6*(p+1)*(r-1))*(1/(n1-1)+1/(n2-1)+1/(n3-1)-1/(n-r))
+     d2=(p-1)*(p+2)/(6*(r-1))*(1/(n1-1)^2+1/(n2-1)^2+1/(n3-1)^2-1/(n-r)^2)
+   }
+   f1=p*(p+1)*(r-1)/2
+   f2=(f1+2)/(d2-d1^2)
+   b=f1/(1-d1-f1/f2)
+   #M/b渐进服从F分布
+   FS=M/b
+   pvalue=pf(FS,f1,f2,lower.tail=FALSE,log.p=FALSE)
+   list("M"=M,"b"=b,"p-value"=pvalue)
+ }

> #test
> cov.test(4,5:8)

$M
[1] 258.4938

$b
[1] 30.40047

$`p-value`
[1] 1.755223e-37
```

检验得到的p-value几乎接近于0，说明4个总体的协方差阵不相同。因此，后续的分析就需要按照协方差阵不同的步骤进行。

Step 3. 计算4个子总体的均值向量和协方差阵

由于距离判别法的判别函数为：

$$V_{ij}(x) = (x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i) - (x-\mu_j)^\top \Sigma_j^{-1}(x-\mu_j)$$

因此，需要计算4个子总体的均值向量和协方差阵。

```
> #计算协方差阵和均值向量
> S1=cov(G1[,5:8])
> S2=cov(G2[,5:8])
> S3=cov(G3[,5:8])
> S4=cov(G4[,5:8])
> m1=colMeans(G1[,5:8])
> m2=colMeans(G2[,5:8])
> m3=colMeans(G3[,5:8])
> m4=colMeans(G4[,5:8])
```

Step 4. 计算所有样本到4个子总体的距离

事实上，判别函数计算的就是每一个样本到各个总体的距离之差。为此，需要分别计算出所有样本到4个子总体的距离。

```
> #计算距离
> d1=c()
> for (i in 1:nrow(data)){
+   dis1=as.matrix(data[i,5:8]-m1)%*%as.matrix(solve(S1))%*%as.matrix(t(data[i,5:8]-m1))
+   d1=c(d1,dis1)
+ }
> d2=c()
> for (i in 1:nrow(data)){
+   dis2=as.matrix(data[i,5:8]-m2)%*%as.matrix(solve(S2))%*%as.matrix(t(data[i,5:8]-m2))
+   d2=c(d2,dis2)
+ }
> d3=c()
> for (i in 1:nrow(data)){
+   dis3=as.matrix(data[i,5:8]-m3)%*%as.matrix(solve(S3))%*%as.matrix(t(data[i,5:8]-m3))
+   d3=c(d3,dis3)
+ }
> d4=c()
> for (i in 1:nrow(data)){
+   dis4=as.matrix(data[i,5:8]-m4)%*%as.matrix(solve(S4))%*%as.matrix(t(data[i,5:8]-m4))
+   d4=c(d4,dis4)
+ }
```

Step 5. 预测及回判

此时，准备工作已经进行完毕。下面，利用判别规则

$$\begin{cases} x \in G_i, & V_{ij}(x) < 0, \forall j \neq i \\ \text{Uncertain}, & \exists V_{ij} = 0 \end{cases}$$

对数据集中的每一个样本进行判别。

```
> D1=c()
> for(i in 1:nrow(data)){
+   if(d1[i]==min(d1[i],d2[i],d3[i],d4[i])){
+     D1=c(D1,1)
+   }
+   if(d2[i]==min(d1[i],d2[i],d3[i],d4[i])){
```

```
+     D1=c(D1,2)
+   }
+   if(d3[i]==min(d1[i],d2[i],d3[i],d4[i])){
+     D1=c(D1,3)
+   }
+   if(d4[i]==min(d1[i],d2[i],d3[i],d4[i])){
+     D1=c(D1,4)
+   }
+ }
```

Step 6. 计算精度

现在，已经通过判别，得到了每一个样本的新等级。那么，判别分析的准确性如何？下面生成误判列联表，并计算判别精度。

```
> ndegree=c()
> for(i in 1:nrow(data)){
+   if(degree[i]==1 || degree[i]==2 || degree[i]==3){
+     ndegree[i]=degree[i]
+   }else{
+     ndegree[i]=4
+   }
+ }
> #误判列联表
> table(ndegree,D1)

       D1
ndegree   1   2   3   4
      1 206   5  28   8
      2 160   5  26  13
      3  42   0   7   2
      4  11   0   0   4

> #判别精度
> sum(diag(prop.table(table(ndegree,D1))))

[1] 0.4294004
```

根据计算结果，大概有43%的样本被判断正确。对于一个难以预测的自然现象，这样的精度尚可接受。但是，在本案例中，评价判别效果不应仅仅关注判别精度；判别精度甚至不应该成为最重要的评价指标。

更合理的评价方式，是观察输出的误判列联表。其中，"D1"代表"判别等级"，"ndegree"代表"真实等级"（合并后）。

在被判入"等级1"的样本中（第1列），实际上也为"等级1"的样本占到将近50%。随着"真实等级"的提高，实际属于"等级2"、"等级3"和"等级4"的样本依次递减。

当"判别等级"是"等级2"（第2列）时，判别精度最高。这可以从2个方面进行说明。一方面，在被判入"等级2"的样本中，实际上也为"等级2"的样本占到50%，这个判对比例在所有的"判别等级"中最高；另一方面，对于那些发生了误判的样本，全部都被判入了"等级1"。通过Table 1，我们得知，"等级2"和"等级1"差别并不是很大——"等级1"代表没有火灾，"等级2"是极度轻微的火灾。因此，即使产生了误判，仍然具有一定的预测作用。

"判别等级"是"等级3"（第3列）和"等级4"（第4列）时，判别效果就比较差了。大部分被判入较高等级的样本，实际上都属于较低的等级。利用这样的判别结果进行火险预警，很有可能出现"过度预警"的效果，从而浪费了大量的人力、财力和物力。

此外，这个判别结果还有一个致命的缺陷，即对"重度火灾"的预测效果不佳。观察误判列联表的第4行，会发现，实际为"等级4"的样本中，大部分都被判为了"等级1"，真正被判对样本只有大概25%。

事实上，在判别过程中，追求"低火险等级"的判别精度往往会牺牲一部分"高火险等级"的判别精度；同样地，追求"高火险等级"的判别精度也会牺牲"低火险等级"的精度。这两个精度存在着矛盾，难以同时兼顾，必须分出主要矛盾和次要矛盾。在这样的取舍中，正所谓"居安思危，思则有备，有备无患，敢以此规"——作为森林火灾的预报和预防，能够对严重的情况进行识别、防备具有更为重要的意义。这是因为，即便低等级火险被误认为高等级火险，也只会促使加强防备而没有过多损失；但是当高等级火险被误认为低等级火险时，却有可能在不经意间造成巨额损失，付出沉痛的代价。

2.7 基于Logistic模型的预报方法改进

2.7.1 想法的来源

基于上述分析，我们需要对判别分析的结果进行改进，从而进一步提高"等级4"火险的判别精度（尽管会损失"等级1"、"等级2"和"等级3"的判别精度）。再一次观察"误判列联表"的第4行，在实际为"等级4"的15个样本中，共有11个样本被判错，并且全部被判入了"等级1"。因此，改进的思路是，对判别分析中被判入"等级1"的全部样本，进行重新分配，使得其中那些实际为"等级4"的样本全部被判入较高等级。

然而，为了找到合适的改进方法，我们首先需要分析，上节中判别分析的效果为什么会不佳。

事实上，判别分析具有3条假设：

（1）每一个判别变量不能是其他判别变量的线性组合；

（2）各组判别变量的协方差矩阵相等；

（3）各组判别变量服从多元正态分布。

其中，假设（1）是为了保证参数估计的稳定性。可以通过计算4个判别变量之间的相关系数，判断是否存在多重共线性。

```
> cor(data.frame(FFMC,DMC,DC,ISI))

          FFMC       DMC        DC       ISI
FFMC 1.0000000 0.3826188 0.3305118 0.5318049
DMC  0.3826188 1.0000000 0.6821916 0.3051278
DC   0.3305118 0.6821916 1.0000000 0.2291542
ISI  0.5318049 0.3051278 0.2291542 1.0000000
```

相关性最强的一组变量，是DC和DMC，它们之间的相关系数为0.68，没有超过0.7。因此，可以认为，判别变量之间并不存在多重共线性，即任意一个判别变量都不是其他变量的线性组合，假设（1）没有违背。

假设（2）则是针对线性判别、二次判别和贝叶斯判别而言的。尽管这里的协方差矩阵不同，距离判别法却能够处理协方差阵异质的情况。因此，假设（2）也不是造成判别效果不佳的原因。

最后，我们需要对假设（3）进行检验，即进行正态分布的检验。

```
> #检验正态分布
> library(mvnormtest)
> Glist=list(G1,G2,G3,G4)
> W=c()
> P.value=c()
> for(i in 1:4){
+   test=mshapiro.test(t(Glist[[i]][,5:8]))
+   W=c(W,test$statistic)
+   P.value=c(P.value,test$p.value)
+ }
> MNtest<-data.frame(W,P.value)
> rownames(MNtest)=paste("G",1:4,sep="")
> MNtest
```

```
           W      P.value
G1 0.3706478 2.112892e-28
G2 0.7228523 3.588740e-18
G3 0.9248119 3.166676e-03
G4 0.6951477 2.206325e-04
```

此处的原假设为H_0：第i个子总体服从多元正态分布。得到的4个p-value全部几乎为0，因此可以认为，数据违背了假设（3），而这或许就是造成判别分析效果不佳的原因。

可见，判别分析的基本假设较为苛刻，因为在自然界中，服从正态分布的数据只是少数。而对于同样具有分类和判别作用的二元Logistic回归模型，则不存在上述种种限制。基于这种考虑，二元Logistic回归模型或许是一种好的改进方法。

2.7.2 基于Logistic回归的"序贯式"算法

根据Logistic回归因变量为0-1变量的特点，借用"序贯式算法"的思想，改进的步骤可叙述如下：

（1）以数据集中的所有观测作为样本，将是否属于"类别1"作为因变量（0-1变量），将$FFMC$、DMC、DC、ISI作为自变量，建立Logistic回归模型，并将该模型命名为"mol1"；

（2）以数据集中的所有观测作为样本，将是否属于"类别2"作为因变量（0-1变量），将$FFMC$、DMC、DC、ISI作为自变量，建立Logistic回归模型，并将该模型命名为"mol2"；

（3）以数据集中的所有观测作为样本，将是否属于"类别3"作为因变量（0-1变量），将$FFMC$、DMC、DC、ISI作为自变量，建立Logistic回归模型，并将该模型命名为"mol3"；

（4）提取出所有先前（在判别分析中）被判入"类别1"的样本，回代入mol1，计算

$$P(y_i = 1|FFMC, DMC, DC, ISI)$$

并且按照

$$\begin{cases} y_i \in degree_1, & P(y_i=1|FFMC,DMC,DC,ISI) > 0.5 \\ y_i \notin degree_1, & P(y_i=1|FFMC,DMC,DC,ISI) \le 0.5 \end{cases}$$

的原则，判断样本是否属于“类别1”；

（5）提取出（4）中没有被判入“类别1”的样本，回代入mol2，计算

$$P(y_i=2|FFMC,DMC,DC,ISI)$$

并且按照

$$\begin{cases} y_i \in degree_2, & P(y_i=2|FFMC,DMC,DC,ISI) > 0.5 \\ y_i \notin degree_2, & P(y_i=2|FFMC,DMC,DC,ISI) \le 0.5 \end{cases}$$

的原则，判断样本是否属于“类别2”；

（6）提取出（5）中没有被判入“类别2”的样本，回代入mol3，计算

$$P(y_i=3|FFMC,DMC,DC,ISI)$$

并且按照

$$\begin{cases} y_i \in degree_3, & P(y_i=3|FFMC,DMC,DC,ISI) > 0.5 \\ y_i \notin degree_3, & P(y_i=3|FFMC,DMC,DC,ISI) \le 0.5 \end{cases}$$

的原则，判断样本是否属于“类别3”；

（7）提取出（6）中没有被判入“类别3”的样本，将它们全部归入“类别4”。

下面，按照上述算法，对原先（判别分析中）被判入“类别1”的419个样本进行重新分配。

```
> data=cbind(data,ndegree,D1)
> dd=subset(data,D1==1)
> Glist=list(subset(data,ndegree==1),
+            subset(data,ndegree==2),
+            subset(data,ndegree==3),
+            subset(data,ndegree==4))
> #---
> #建立mol1
> A=Glist[[1]]
> B=rbind(Glist[[2]],Glist[[3]],Glist[[4]])
> A$ndegree=1
> B$ndegree=0
> mol.1=glm(ndegree~FFMC+DMC+DC+ISI,
+        family = binomial(link = logit),data=rbind(A,B))
> #---
> #建立mol2
> A=Glist[[2]]
> B=rbind(Glist[[3]],Glist[[4]])
> A$ndegree=1
```

```
> B$ndegree=0
> mol.2=glm(ndegree~FFMC+DMC+DC+ISI,
+         family = binomial(link = logit),data=rbind(A,B))
> #---
> #建立mol3
> A=Glist[[3]]
> B=rbind(Glist[[4]])
> A$ndegree=1
> B$ndegree=0
> mol.3=glm(ndegree~FFMC+DMC+DC+ISI,
+         family = binomial(link = logit),data=rbind(A,B))
> #---
> #进行判别
> y=c()
> for(k in 1:nrow(dd)){
+   logit=predict(mol.1,data.frame(FFMC=dd$FFMC[k],DMC=dd$DMC[k],
+                                  DC=dd$DC[k],ISI=dd$ISI[k]))
+   pro=exp(logit)/(1+exp(logit))
+   if(pro>0.5){
+     y=c(y,1)
+   }else{
+     logit=predict(mol.2,data.frame(FFMC=dd$FFMC[k],DMC=dd$DMC[k],
+                                    DC=dd$DC[k],ISI=dd$ISI[k]))
+     pro=exp(logit)/(1+exp(logit))
+     if(pro>0.5){
+       y=c(y,2)
+     }else{
+       logit=predict(mol.3,data.frame(FFMC=dd$FFMC[k],DMC=dd$DMC[k],
+                                      DC=dd$DC[k],ISI=dd$ISI[k]))
+       pro=exp(logit)/(1+exp(logit))
+       if(pro>0.5){
+         y=c(y,3)
+       }else{
+       y=c(y,4)
+       }
+     }
+   }
+ }
> dd$D1<-y
> ddd=subset(data,D1!=1)
> N=rbind(dd,ddd)
```

重新分配后，得到了新的的“误判列联表”。

```
> t1=table(N$ndegree,N$D1)
> #误判列联表
> t1
```

	1	2	3	4
1	70	141	28	8
2	41	124	26	13
3	14	28	7	2
4	2	9	0	4

```
> #判别精度
> sum(diag(prop.table(t1)))
```

```
[1] 0.3965184
```

然而，在实际属于“类别4”的样本中，仍然有2个被判为“类别1”。这说明，在本案例的特殊情况下，在计算得到$P(y_i = k|FFMC, DMC, DC, ISI)$之后，需要对判别原则进行进一步改进。

2.7.3 “阀门”临界值的调整

上述的判别规则均以

$$\begin{cases} y_i \in degree_k, & P(y_i = k|FFMC, DMC, DC, ISI) > 0.5 \\ y_i \notin degree_k, & P(y_i = k|FFMC, DMC, DC, ISI) \leq 0.5 \end{cases}$$

为原则，其中的临界值为0.5。然而，如果希望实际属于“等级4”的样本全部被判入较高等级（“等级3”或“等级4”）中，0.5的临界值显然有些低了。

形象地看，在进行判别时，我们一共使用了3个“阀门”，即3个临界值，分别用来控制样本是否属于“类别1”，是否属于“类别2”，以及是否属于“类别3”，并将它们分别命名为“阀门1”、“阀门2”和“阀门3”。通过压低某一类的“阀门”临界值，可以使得更多的样本被判入此类；而抬高“阀门”临界值，就会使得样本难以被判入此类别，从而被判入其后续的类别。

我们所希望的最佳结果，是所有属于“等级4”的样本全部被判对。但是这同样也会使过多属于较低等级（“等级1”或“等级2”）的样本被判入“等级4”，从而出现过度“警报”的情况，矫枉过正。因此，退而求其次，可以接受相当一部分实际属于“等级4”的样本被判入“等级3”；但绝不可被判入“等级2”或“等级1”。

“阀门”临界值调整的原则和顺序如下：

（1）所有“阀门”调整的精度为4位小数；

（2）以0.5为起点，抬高“阀门1”，每次增加0.0001，直至所有实际属于“类别4”的样本不再被判入“类别1”；

（3）以0.5为起点，抬高“阀门2”，每次增加0.0001，直至所有实际属于“类别4”的样本不再被判入“类别2”；

（4）以0.5为起点，抬高“阀门3”，每次增加0.0001，直至所有实际属于“类别4”的样本，被判入“类别4”的数量多于被判入“类别3”的数量。

调整过程如下所示：

Table 2: “阀门1”（误判列联表第一列）的调整过程

临界值	实际“类别1”样本数	实际“类别2”样本数	实际“类别3”样本数	实际“类别4”样本数
0.5000	70	41	14	2
0.5003	69	41	14	1
0.5037	68	41	14	0

Table 3: “阀门2”（误判列联表第二列）的调整过程

临界值	实际“类别1”样本数	实际“类别2”样本数	实际“类别3”样本数	实际“类别4”样本数
0.5000	143	124	28	11
0.7000	139	121	28	10
0.7153	131	141	25	9
0.7204	129	109	25	8
0.7382	98	88	18	5
0.7393	90	80	16	4
0.7399	87	77	16	3
0.7486	70	67	14	2
0.7663	40	50	9	1
0.7982	20	29	5	0

Table 4: “阀门3”（误判列联表第三列）的调整过程

临界值	实际“类别1”样本数	实际“类别2”样本数	实际“类别3”样本数	实际“类别4”样本数
0.5000	142	110	28	10
0.6228	135	110	24	9
0.6858	124	100	21	8
0.7249	108	85	18	7

最终，得到如下结果：

```
> y=c()
> for(k in 1:nrow(dd)){
+   logit=predict(mol.1,data.frame(FFMC=dd$FFMC[k],DMC=dd$DMC[k],
+                              DC=dd$DC[k],ISI=dd$ISI[k]))
+   pro=exp(logit)/(1+exp(logit))
+   if(pro>0.5037){
+     y=c(y,1)
+   }else{
+     logit=predict(mol.2,data.frame(FFMC=dd$FFMC[k],DMC=dd$DMC[k],
+                                DC=dd$DC[k],ISI=dd$ISI[k]))
+     pro=exp(logit)/(1+exp(logit))
+     if(pro>0.7982){
+       y=c(y,2)
+     }else{
+       logit=predict(mol.3,data.frame(FFMC=dd$FFMC[k],DMC=dd$DMC[k],
+                                  DC=dd$DC[k],ISI=dd$ISI[k]))
+       pro=exp(logit)/(1+exp(logit))
+       if(pro>0.7249){
+         y=c(y,3)
+       }else{
+       y=c(y,4)
+       }
+     }
+   }
```

```
+ }
> dd$D1<-y
> ddd=subset(data,D1!=1)
> N=rbind(dd,ddd)
> t1=table(N$ndegree,N$D1)
> #"误判列联表"
> t1

      1   2   3   4
  1  68  20 108  51
  2  41  29  85  49
  3  14   5  18  14
  4   0   0   7   8

> #"计算精度"
> sum(diag(prop.table(t1)))

[1] 0.237911
```

2.7.4 结果的评价与特点

此时，"阀门1"被调整至0.5037，"阀门2"被调整至0.7982，"阀门3"被调整至0.7249。15个实际为"类别4"的样本中，7个被判入了"类别3"，8个被判入了"类别4"。尽管判别精度下降到了23.79%，但是过分地关注精度已经意义不大。这是因为，在后续的分析中，对于火灾的"预测"，已经转化为对火灾"发生模式"的预测，这种模式包括火灾发生月份、"小规模火灾"发生频率以及发生"重度火灾"的可能性；而对火灾"发生模式"的判断比单纯地预测等级更加有用。

观察现在的"误判列联表"，不难发现，我们的目的已经达到，所有实际为"等级4"的样本都被判入了较高的等级，并且，其中被判入"等级4"的样本多于被判入"等级3"的样本。然而，也有相当多的低等级样本被判入高等级。那么，这样的结果会对最终的预报和预防产生多大的影响呢？

结合2.4节中我们对月份和火灾等级的分析，此处不妨将每一个判别等级按月份进行频数统计。

```
> K1=subset(N,D1==1)
> addmargins(table(K1$ndegree,K1$month),2)

    apr aug dec feb jan jul jun mar may nov oct sep Sum
  1   5   2   0  10   2   3   8  35   1   1   0   1  68
  2   3   0   6   8   0   3   6  14   0   0   0   1  41
  3   1   0   3   2   0   1   1   5   1   0   0   0  14
```

在被判入"等级1"的样本中，几乎各月都有险情，甚至大部分发生在春秋季节。总体以"等级1"和"等级2"为主，呈现小范围过火的特点。

```
> K2=subset(N,D1==2)
> addmargins(table(K2$ndegree,K2$month),2)

    apr aug dec feb jan jul jun mar may nov oct sep Sum
  1   0  17   0   0   0   0   0   0   0   0   0   3  20
  2   0  25   0   0   0   0   0   0   0   0   0   4  29
  3   0   3   0   0   0   0   0   0   0   0   0   2   5
```

在被判入"等级2"的样本中，只有8月和9月出现了险情。总体仍然以小规模火灾为主。但是，由于8月和9月本就是需要严加防范的月份，这样的判别结果几乎是无用的。

```
> K3=subset(N,D1==3)
> addmargins(table(K3$ndegree,K3$month),2)

    apr aug dec feb jan jul jun mar may nov oct sep Sum
  1   0  42   0   0   0  11   1   0   0   0  10  44 108
  2   0  37   0   0   0  10   1   0   0   0   3  34  85
  3   0   8   0   0   0   0   0   0   0   0   2   8  18
  4   0   1   0   0   0   2   0   0   0   0   0   4   7
```

在被判入"等级3"的样本中，6月到10月均出现了险情。此时，集中于夏季的低等级、小规模的灾情频数明显上升，并且伴随着较多的"等级4"重度火灾。这意味着，一旦出现被判入"等级3"的情况，一般的防范措施或许就不够了。

```
> K4=subset(N,D1==4)
> addmargins(table(K4$ndegree,K4$month),2)

    apr aug dec feb jan jul jun mar may nov oct sep Sum
  1   0  24   0   0   0   0   0   0   0   0   0  27  51
  2   0  17   0   0   0   2   0   0   0   0   0  30  49
  3   0   4   0   0   0   0   0   0   0   0   0  10  14
  4   0   4   0   0   0   0   0   0   0   0   0   4   8
```

在被判入"等级4"的样本中，只有7月到9月出现了险情。此时，不仅集中在这几个月的小规模灾情较多，而且"等级4"重度火灾的比例和频数都较高。

综上所述，尽管有相当多的低等级样本被判入高等级，但是这些误判的样本**恰好全部发生在高等级火灾频发的月份**，而这也恰好与7月到9月间"小规模火灾频繁发生"且"出现重度火灾"的特点相呼应。也就是说，这样的误判并不会造成"过度预警"，由于误判而增加的防范措施反而会对"重度火灾"的预防与救援大有裨益。

3 结论

3.1 预警预防方法的建立

在上述分析中，我们利用基于偏态和"上极端值"的方法，对过火面积$area$进行了聚类，将其划分为若干个等级；采用典型相关分析，揭示了FFMC指数、DMC指数、DC指数和ISI指数与各个气象指标之间的密切关系，并指出"FWI系统指数"更加适合用作多元建模；利用对应分析，得出了7月至9月为火灾高发月的结论，在此期间应着重防范；利用对数线性模型，说明了在休息日和节日期间防范轻微火灾的现实意义；最后，利用判别分析和二元Logistic模型，建立并改进了火险等级的判别方法。此外，在2.7.4节中，针对最终得到的"误判列联表"，我们还给出了各个"判别等级"的特点。

事实上，每一个"判别等级"都可以代表一种特定的"火灾发生"模式。结合2.1至2.7节中的分析结果，以及不同的"模式"，我们可以构建预警体系，并对Montesinho自然公园的森林火灾防范工作提出以下建议：

（1）将7月至9月作为重点防范月，在此期间的防范措施，要达到能够应对至少80公顷（“等级3”）过火面积的程度；

（2）根据收集到的气象数据，每日计算“FWI系统指数”，代入2.6节中得到的判别模型，计算“判别等级”；

（3）如果（2）中得到的“判别等级”为“等级1”，则将这部分数据代入2.7.3节中得到的二元Logistic“序贯”模型，重新划分等级；如果（2）中得到的“判别等级”为其他等级，则不进行更改；

（4）如果最终的“判别等级”为“等级2”，那么这样的判别结果是无用的；因为这种情况往往只发生在8月和9月，而8月至9月间的防范措施完全能够应对“判别等级”为“等级2”时的最严重情况。此时发出“蓝色预警”，意味着无需增加防范程度，保持现状即可；

（5）如果最终的“判别等级”为“等级1”，那么“火灾模式”为春秋季节的“小规模火灾”。此时发出“黄色预警”，意味着需要重视节日期间、休息日期间的火灾风险，对小规模火灾提高警惕；

（6）如果最终的“判别等级”为“等级3”，那么“火灾模式”表现为夏季“小规模火灾”的骤增以及“重度火灾”的出现。此时发出“红色预警”，意味着在现行防范措施的基础上，需要加强防备，务必重视“重度火灾”发生的可能性；

（7）如果最终的“判别等级”为“等级4”，那么极有可能会发生最严重的“超级火灾”。此时发出“黑色预警”，即使全员戒备也不为过。

事实上，“蓝色预警”和“黑色预警”一般只发生在7月至9月，“红色预警”一般也只发生在6月至10月。真正对秋冬季节有意义的，是“黄色预警”。秋冬季节火灾较少，往往防备松懈，一旦收到“黄色预警”，就必须引起重视，在休息日和节日期间尤其如此。

3.2 模型的评价与讨论

在2.2节过火面积$area$的聚类分析中，尝试了一种基于“极端值”的聚类方法。事实上，这种方法具有普遍的意义——只要更改$(max - Q_3)/IQR$的临界值，就可以人为地控制聚类数目，从而为单变量的聚类提供新的思路。

此外，本文采用的分析方法，完全属于多元统计分析的领域。由于精度不高，因此在很多方面进行了“妥协”，并且模型的改进、得到的结论几乎完全囿于特定数据，仅仅对Montesinho国家公园的火灾预警和防范具有一定意义。对于其他地区的森林火灾，这样的模型是否有效，仍然值得商榷。

事实上，在火灾预测方面还有不少精度更高的方法，包括支持向量机、决策树、神经网络、随机森林和粒子蚁群算法等。利用上述模型对预测方法加以改进，或许会得到更好的结果。

4 附录

1. 线性判别法的代码

```
> ld=lda(ndegree~FFMC+DMC+DC+ISI,data)
> new=predict(ld)
> t2=table(ndegree,new$class)
> #误判列联表
> t2

ndegree   1   2   3   4
      1 213  34   0   0
      2 166  38   0   0
      3  42   9   0   0
      4  15   0   0   0

> #判别精度
> sum(diag(prop.table(t2)))

[1] 0.4854932
```

2. 二次判别法的代码

```
> qd=qda(ndegree~FFMC+DMC+DC+ISI,data)
> new=predict(qd)
> t3=table(ndegree,new$class)
> #误判列联表
> t3

ndegree   1   2   3   4
      1  59 183   0   5
      2  39 156   0   9
      3  11  39   0   1
      4   0  13   0   2

> #判别精度
> sum(diag(prop.table(t3)))

[1] 0.4197292
```

References

[1] 信晓颖，江洪等, “加拿大森林火险气候指数系统（FWI）的原理及应用,” 浙江农林大学学报, vol. 28, no. 2, pp. 314 – 318, 2011.

[2] Cortez and Morais, “A data mining approach to predict forest fires using meteorological data,” in *New Trends in Artificial Intelligence, Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence*, (Guimarães, Portugal), pp. 512–523, December 2007.