第 19 卷 第 2 期
2010 年 4 月

火 灾 科 学
FIRE SAFETY SCIENCE

Vol. 19, No. 2
Apr. 2010

文章编号：1004-5309(2010)-0082-07

一种基于小样本数据信息扩散的重大火灾频度估算方法

李炳华[1]，朱霁平[1*]，小出治[2]，彭 晨[1]

（1. 中国科学技术大学 火灾科学国家重点实验室，安徽 合肥，230026；
2. 日本东京大学 工学部都市工学系，日本）

摘要：重大火灾的发生是小概率事件，其历史数据属于小样本统计数据。在应用频率直方图法对小样本数据进行处理时，分析结果对直方图的起点、区间步长以及样本中的奇异数据十分敏感。采用信息扩散方法，对日本 1995-2008 年各年重大火灾次数统计样本进行分析，对比了不同扩散条件下的扩散结果，结果表明信息扩散方法具有很好的稳定性和一致性。基于信息扩散的结果，计算了以年为周期的重大火灾发生次数的超越概率分布，建立了一种重大火灾频度估算方法。

关键词：信息扩散；重大火灾；小样本；数理统计
中图分类号：X928.7　　文献标识码：A

0 引言

火灾风险评估是火灾科学的新兴研究领域[1]。由于火灾具有确定和随机双重特性，运用概率论和数理统计等数学方法，对灾害历史统计数据进行分析，是一种简单实用的风险分析方法，不要求对致灾原理及其动力学演化过程有深入的掌握。统计分析有非常悠久的历史，对其估计方法的研究已经非常的完善和深入[2]。一个统计结果是否有效，取决于合理的假设估计模型和足够的样本两个条件[3]。然而，在很多实际工程问题中很难找到足够大的样本。火灾风险评估研究中存在许多非完备性小样本分析问题[4]，例如重大火灾是小概率事件，相关历史统计数据属于小样本。对于这类小样本统计数据来说，要预先给出一个合理的估计模型进行参数估计，是十分困难的。

宋卫国[5]提出可以根据火灾频率的幂律分布规律，利用小火灾发生频率来估算大火灾发生频率，然而由于数据样本的缺失，用该方法来预测重大火灾风险的可行性还有待讨论。以“信息分配”和“信息扩散”为核心的模糊信息优化处理技术是有中国学者独立提出和发展的一门新兴数据处理技术[6]。信息扩散是为了弥补样本信息不足而考虑优化利用样本模糊信息的一种集值化处理方法。信息扩散在灾害风险评估方面已经得到了初步的应用[7,8]。景国勋等基于信息扩散理论，根据河南省各市级行政区历年火灾损失率对未来火灾损失率进行预测[9]。张继权等应用信息扩散技术对吉林省 11 年草原历史火灾数据进行处理，对草原火灾次数、过火面积、经济损失的风险进行计算，定量地评价了吉林省草原火灾风险[10]。上述研究直接采用了信息扩散方法对火灾统计数据进行处理，对方法的适用性并没有进行讨论分析。同时，上述研究直接采用信息扩散的结果来预测火灾风险，当小样本数据中存在个别

收稿日期：2010-02-13；修改日期：2010-03-25
基金项目：国家科技支撑计划项目 2006BAD04B05-04；国家林业公益性行业科研专项重点项目 200704027。
作者简介：李炳华，男，中国科学技术大学火灾科学国家重点实验室硕士研究生，从事火灾风险研究。
通讯作者：朱霁平，E-mail：jpzhu@ustc.edu.cn.

奇异数据时，预测结果会有较大的偏差。

本文选用日本 1995-2008 年各年重大火灾次数小样本统计数据，首先采用传统的频率直方图法进行表征，发现该方法对直方图的起点、区间步长以及样本中的奇异数据十分敏感。然后采用信息扩散方法对样本数据进行处理，分别讨论论域起始点、论域步长及扩散函数对扩散结果的影响；在此基础上，建立了一种基于信息扩散的重大火灾频度估算方法。

1　数据样本

本文选用日本 1995 年～2008 年重大火灾次数统计数据进行分析，见表 1。

表 1　日本 1995-2008 年重大火灾次数

Table 1　The numbers of major fires between 1995 and 2008 in Japan

1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
93	50	57	55	52	59	60	61	71	61	44	48	62	57

图 1 表示日本 1995-2008 年重大火灾次数随年份的变化情况。从中可以看出，每年重大火灾发生次数并没有明显的变化趋势，很难从时间维度上做出合理的预测。其中，1995 年由于阪神地震导致日本重大火灾次数明显高出其他年份，该数据属于小样本数据集中的奇异数据。

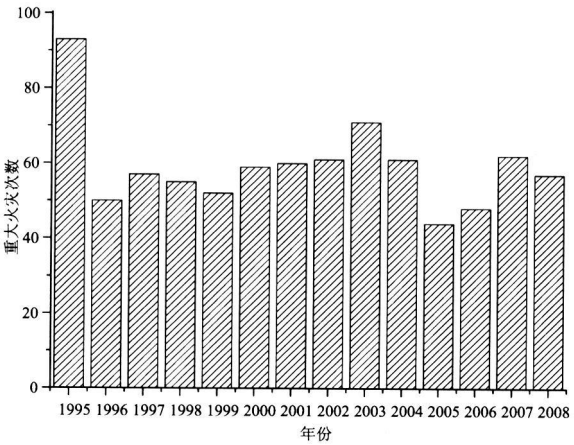

图 1　日本 1995-2008 年重大火灾次数

Fig. 1　The numbers of major fires between 1995 and 2008 in Japan

2　数据处理方法

2.1　频率直方图处理法

图 2 中四个柱状图表示了分别取不同区间长度和不同起始统计点时频数分布情况。首先由于样本数据集中存在奇异数据，每个频率直方图中都出现了空白矩形。然后通过两两比较图 2 中四张直方图的形状和变化趋势，可以看出直方图起点和区间步长对直方图的形状和变化趋势有较大的影响。说明在处理小样本数据时，直方图方法的稳定性较差。直接采用频率直方图处理，很难对数据进行有效的拟合。

2.2　信息扩散方法

信息扩散是为了弥补样本信息不足而考虑优化利用样本模糊信息的一种集值化处理方法[3]。信息扩散处理方法的基本步骤，首先把原始样本点分散的信息逐一映射到等步长的一组控制点上，由扩散函数确定各个控制点获得的信息量，然后对控制点各自进行信息量累加，得到新的样本集。以本文处理的数据为例，原始样本输入向量为 $X=(93, 50, 57, 55, 52, 59, 60, 61, 71, 61, 44, 48, 62, 57)$，样本数量 $m=14$。转换为控制点论域空间为 $U=[u_1, u_2, \cdots, u_n]$，其中 u_1 是论域的起始控制点，$u_n=u_1+(n-1)\Delta$ 是结束控制点，Δ 是论域的步长，n 是控制点个数。各控制点的累积信息量可根据下式计算。

$$g(u_i)=\sum_{j=1}^{j=m}\frac{\mu(u_i, x_j)}{\sum_{i=1}^{i=n}\mu(u_i, x_j)} \tag{1}$$

其中：$g(u_i)$ 是各控制点分配到的信息量，$\mu(u_i, x_j)$ 为不同形式的扩散函数。

图 3 表示在不同扩散条件下得到的各控制点累积信息量。可以看出各直方图的变化趋势非常相似，表明在不同扩散条件下得到的扩散结果具有较好的一致性。

控制点的累积信息量体现的是以控制点为中心，步长为长度的区间信息总量。采用以下公式进一步归一化处理得到各控制点的概率值：

$$p(u_i)=\frac{g(u_i)}{\Delta\cdot m} \tag{2}$$

其中：$p(u_i)$ 是控制点概率值，$g(u_i)$ 由(1)式计算，Δ 为控步长，m 为原始样本总数。

为了进一步分析不同扩散条件对扩散结果的影响，下面利用 MATLAB 曲线拟合工具对(2)式求得的控制点概率值进行曲线拟合，通过对比拟合参数，讨论了起始控制点、步长及扩散函数三种扩散条件

对结果的影响。

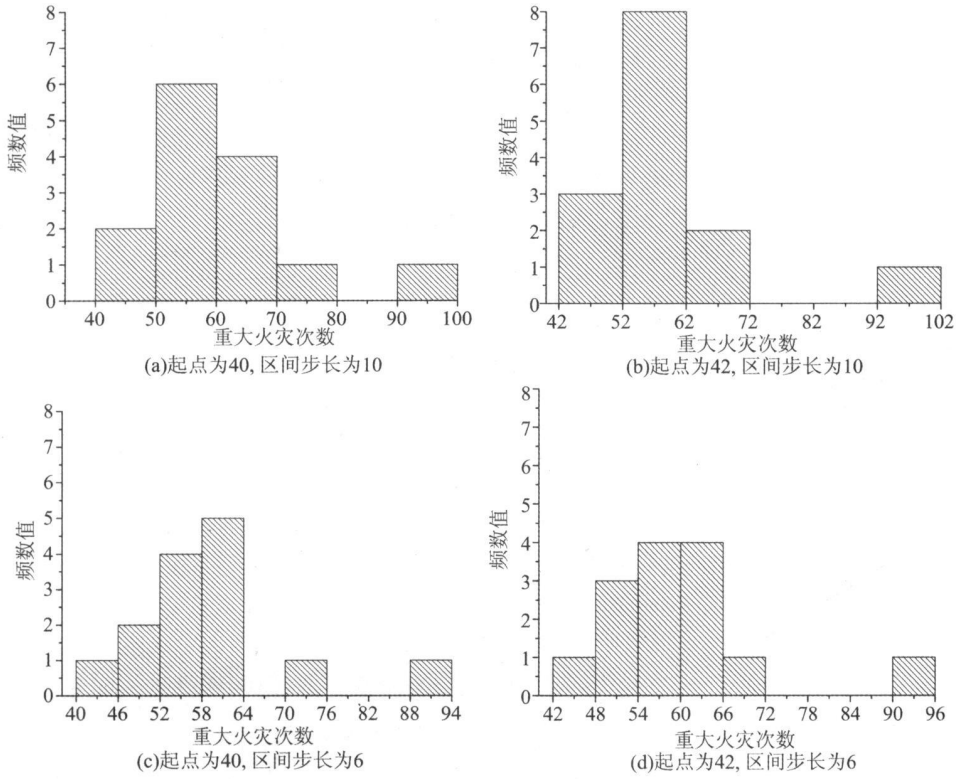

图 2 不同直方图起点和区间步长时的频率直方图

Fig. 2 Frequency histograms under different lengths of interval or starting points

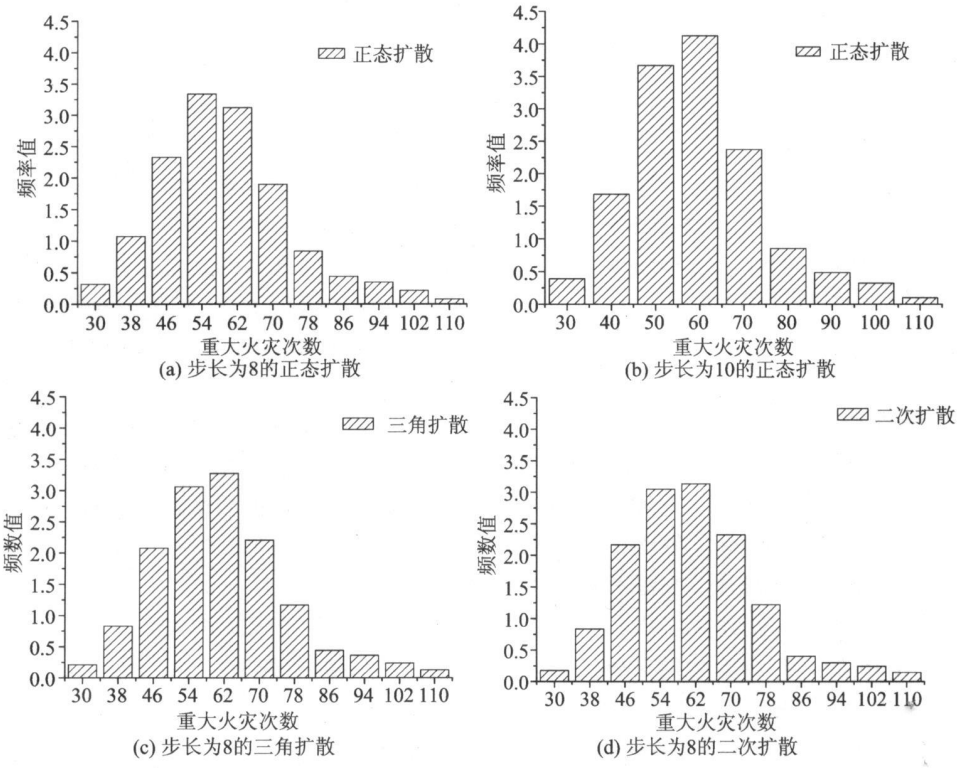

图 3 不同扩散函数及论域步长条件下的各控制点的累积信息量

Fig. 3 Total Distribution Information under different diffusion functions or universes step sizes

1）　起始控制点对扩散结果的影响

起始控制点分别取为 30、32、34，固定论域步长为 8，选定扩散函数为正态扩散，分别计算三种情况下各个控制点的点概率值，然后进行曲线拟合。表 2 为高斯拟合结果，方程形式为式(3)。拟合方程相关系数 R^2 都在0.985以上，说明各控制点的概率分布符合高斯分布，即正态分布。并且三条拟合曲线的方程系数最大相差为 0.7%。从图 4 也可以看出，三条拟合曲线基本重合，说明不同起始控制点条件下的扩散结果具有相当好的一致性。

$$f(x) = a \cdot \exp\left(-\frac{(x-b)^2}{c^2}\right) \tag{3}$$

表 **2**　不同起始控制点时高斯拟合方程系数

Table 2　Equation coefficients of gauss curve fitting with different first controlling point

	起始控制点	a	b	c	R^2
拟合方程 1	30	0.03022	56.93	17.85	0.9863
拟合方程 2	32	0.03032	56.89	17.83	0.9866
拟合方程 3	34	0.03044	56.84	17.85	0.9870

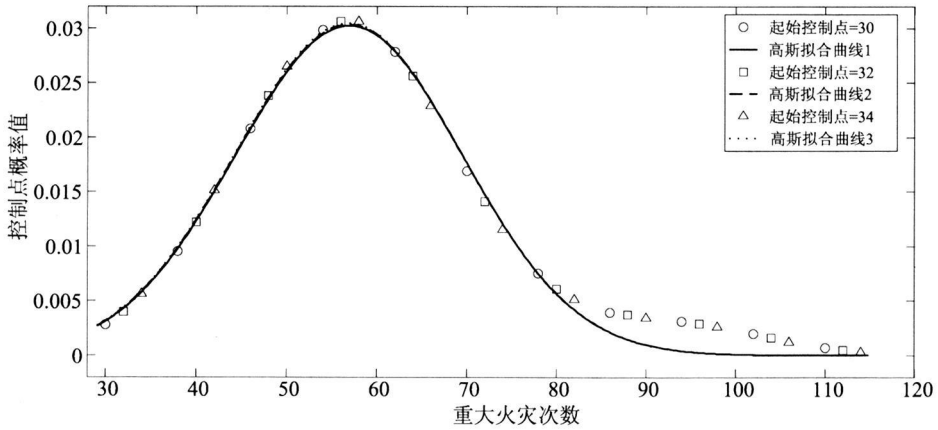

图 **4**　起始控制点不同时概率密度曲线拟合

Fig. 4　Gauss curve fitting with different first controlling points

2）　步长对扩散结果的影响

分别取论域的区间步长为 1～20，起始控制点统一为 30，选定扩散函数为正态扩散，进行信息扩散处理后得到各拟合方程的系数如表 3 所示。可以看出拟合方程相关系数 R^2 都大于 0.98。当步长值在 15 以内时，拟合方程系数的最大相差为 0.9%，此时拟合曲线基本重合。而当步长增大至 16 及以上时，拟合方程系数的最大相差扩大为 9.8%，拟合方程参数有较大偏差。图 5 为步长分别取 1、8、15、20 时的拟合曲线，只有步长为 20 的拟合曲线有一定的偏差。由此证明，论域步长的取值对扩散结果有一定影响，但是在一个比较宽的取值范围内其影响很小。

表 **3**　不同论域步长时高斯拟合方程系数

Table 3 Equation coefficients of gauss curve fitting under different universe step sizes

步长	a	b	c	R^2	步长	a	b	c	R^2
1	0.03034	56.86	17.97	0.9846	11	0.03017	56.94	17.83	0.986
2	0.03031	56.88	17.97	0.9849	12	0.03022	56.93	17.72	0.988
3	0.03027	56.89	17.97	0.9841	13	0.03029	56.94	17.7	0.9864
4	0.03032	56.89	17.86	0.9852	14	0.03045	56.94	17.52	0.9876
5	0.03028	56.92	17.85	0.9855	15	0.03044	56.89	17.45	0.9883
6	0.03023	56.92	17.89	0.9851	16	0.02989	57.07	18.1	0.9876
7	0.03025	56.94	17.83	0.9848	17	0.02942	57.13	18.58	0.9898

8	0.03022	56.93	17.85	0.9863	18	0.02885	57.11	19.17	0.9917
9	0.03024	56.93	17.8	0.9867	19	0.02862	56.91	19.36	0.9883
10	0.03035	56.93	17.64	0.9861	20	0.0292	56.66	18.69	0.986

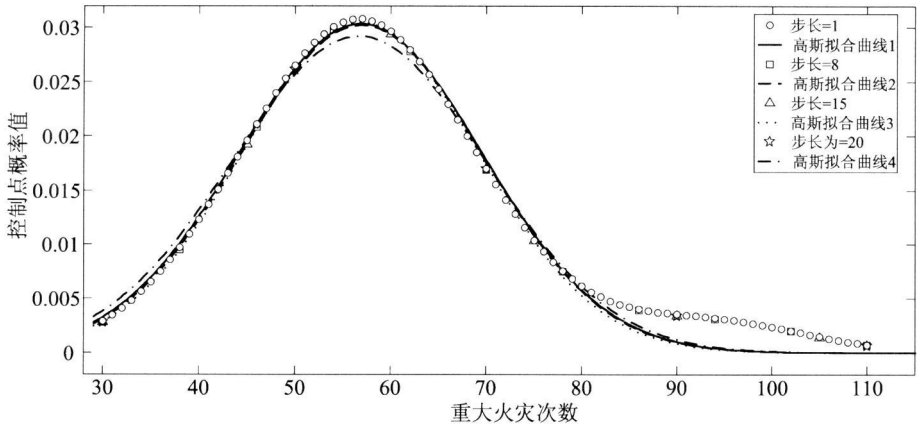

图 5 控制点个数不同时概率密度曲线拟合

Fig. 5 Gauss curve fitting under different step sizes

3) 扩散函数对扩散结果的影响

分别采用正态扩散、三角扩散、和二次扩散[11]三种形式的扩散函数，起始控制点选为 30，步长选为 8，拟合方程系数如表 4。从中可以看出，无论采用哪种扩散函数，正态拟合相关系数都超过 0.98。由此可见，采用不同的扩散函数，并不会影响原始数据的内在规律。同时，不同扩散函数对拟合方程系数存在一定的影响。

表 4 不同扩散函数时高斯拟合方程系数

Table 4 Equation coefficients of gauss curve fitting under different diffusion functions

	扩散函数	a	b	c	R^2
拟合方程 1	正态	0.03022	56.93	17.85	0.9863
拟合方程 2	三角	0.02954	59.00	18.41	0.9839
拟合方程 3	二次	0.02692	59.17	20.64	0.9858

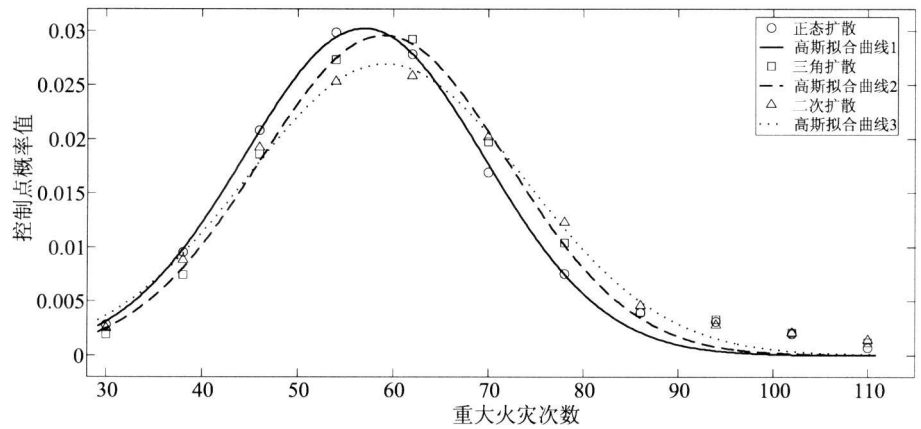

图 6 采用不同扩散函数时概率密度曲线拟合

Fig. 6 Gauss curve fitting under different diffusion functions

3 重大火灾频度估算方法

3.1 基于信息扩散的重大火灾频度估算方法

本文第 3 节证明了信息扩散方法在不同扩散条件下的扩散结果具有很好的稳定性和一致性，由此可以建立一种重大火灾频度大小的估算方法。

一年内重大火灾次数为 x，其论域为 $U=[a, b]$，即 $x \in [a, b]$。a, b 分别为重大火灾次数的下限

和上限。在一定扩散条件下，依据公式(1)、(2)计算各控制点概率值，并进行曲线拟合得到概率密度函数 $f(x)$，$f(x)$表示一年内发生 x 次重大火灾的概率值，累积概率值 $F(x)$表示重大火灾次数小于等于 x 的概率值，

$$F(x)=\sum_{u=a}^{x} f(u) \tag{4}$$

相应的超越概率值 $G(x)$表示重大火灾发生次数大于 x 的概率值，

$$G(x)=1-F(x) \tag{5}$$

定义

$$Z(x)=\frac{1}{G(x)} \tag{6}$$

表示重大火灾发生次数超过 x 的周期(年)，以此评估重大火灾的风险程度。

3.2　实例

以本文第 2 节所述数据样本为研究对象。设重大火灾风险论域为[0, 110]，选择起始控制点为 30，步长为 8、扩散函数为正态扩散。计算各控制点概率值后拟合方程为：

$$f(x)=0.03022\exp\left(-\frac{(x-56.93)^2}{17.85^2}\right) \tag{7}$$

根据公式(4)、(5)、(6)，可以做出如图 7 所示的累积概率密度曲线。从图中可以看出，重大火灾次数小于等于 86 的累积概率 F(86)为 0.9905，相应的 Z(86)=104.8，即 86 次/年的重大火灾次数约为百年一遇；同样可以得出，大约十年就有一年的重大火灾次数超过 73 次；五年中会有一年的重大火灾次数超过 67 次。

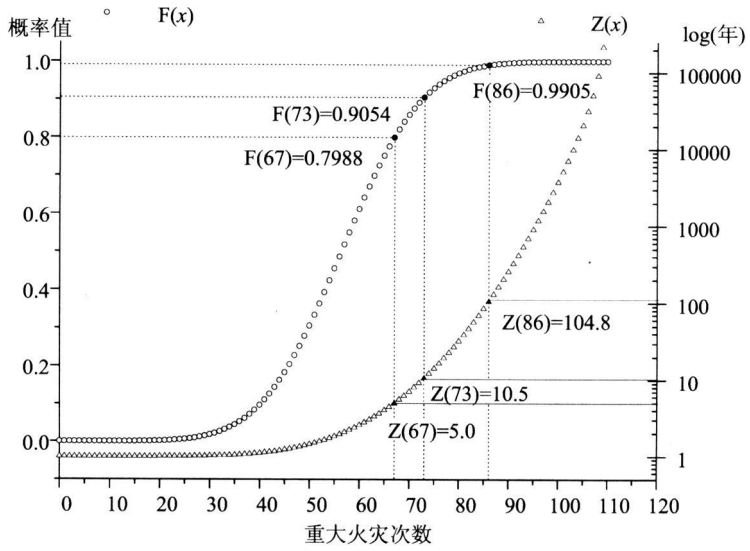

图 7　累积概率密度曲线

Fig. 7　Cumulative probability density curve

4　结论

以日本 1995-2008 年各年重大火灾次数统计样本为例，采用信息扩散方法对数据进行处理，发现在不同扩散条件下控制点累积信息量直方图具有较好的连续性，相对传统的频率直方图法更为优越。同时，在不同扩散条件下，扩散结果具有很好的稳定性和一致性。具体为：不同起始控制点对扩散结果没有影响；不同步长有一定的影响，但可以找到一个比较宽的取值范围，在此范围内可以忽略步长的影响；采用不同的内部扩散函数不会影响样本内在的变化规律，但对拟合参数有一定影响。在此基础上，提出了一种基于信息扩散的重大火灾频度的估算方法，可以在小样本条件下，对重大火灾发生次数的频度给出合理的预测结果。

参考文献

[1] 范维澄, 刘乃安. 中国火灾科学基础研究进展与展望[J]. 中国科学技术大学学报, 2006, 36(1): 1~8.

[2] 王凯平, 张辉. 试论模糊回归与统计回归[J]. 统计与决策, 2009, 12: 30~32.

[3] 黄崇福. 自然灾害风险评估方法理论与应用[M]. 北京: 科学出版社, 2005: 81~82.

[4] 汪金辉, 陆守香. 建筑火灾中人员安全疏散的可靠概率分析模型[J]. 中国科学技术大学学报, 2006, 36(1): 116~118.

[5] 宋卫国, 王健. 火灾系统的复杂性与可持续防治[J]. 科技导报, 2004, 8: 15~18.

[6] 黄崇福, 张俊香, 刘静, 等. 模糊信息优化处理技术应用简介[J]. 信息与控制, 2004, 33(1): 61~66.

[7] 冯利华, 程归燕. 基于信息扩散理论的地震风险评估[J]. 地震学刊, 2000, 20(1): 19~22.

[8] L. H. Feng, C. F. Huang. A Risk Assessment Model of Water Shortage Based on Information Diffusion Technology and its Application in Analyzing Carrying Capacity of Water Resources[J]. Water Resource Manage, 2008, 22: 621~633.

[9] 景国勋, 王卫敏, 刘秋菊, 等. 信息扩散理论在河南省火灾风险评估中的应用[J]. 工业安全与环保, 2008, 34(05): 62~64.

[10] 张继权, 刘兴朋. 基于信息扩散理论的吉林省草原火灾风险评价[J]. 干旱区地理, 2007, 30(4): 590~594.

[11] 陈志芬, 黄崇福, 张俊香, 等. 基于扩散函数的内集—外集模型[J]. 模糊系统与数学, 2006, 20(1): 42~48.

A Method of serious fire risk prediction based on small sample information diffusion

LI Bing-hua[1], ZHU Ji-ping[1], Osamu Koide[2], PENG Chen[1]

(1. State Key Laboratory of Fire Science, USTC, Anhui Hefei, 230026, China;
2. Department of Urban Engineering, University of Tokyo, Japan)

Abstract: The serious and extremely serious fires are events with low probabilities, and the statistics data of them are of small sample. The common frequency histogram method is not applicable to analyze the small sample data, because it is sensitive to the analysis parameters such as the starting point and the interval step. In this paper, based on the method of information diffusion, the statistics data of the serious fires in Japan between 1995 and 2008 are analyzed. The results show that the information diffusion method has good stability and consistency. It has been proved that the influences of changing the starting point, the interval step, or the diffusion function can be ignored. Based on the results of information diffusion, the probability of the number of the serious fires per annum can be fitted with the Gaussian distribution.

Keyword: Information Diffusion; Serious Fire; Small Sample; Statistics