大 气 科 学
1993 年增刊 SCIENTIA ATMOSPHERICA SINICA Sup., 1993

历史旱涝等级空间相关列联概率插值方法*

王 其 冬 项 静 恬
（中国科学院自然资源综合考察委员会，北京 100101） （中国科学院应用数学研究所）

提 要

本文选取自1736 至 1979 年华北、江淮、华东和中南等地区具有代表性的 13 个站的历史旱涝等级资料[1]，用空间相关列联概率方法，对各站某些年代的缺值进行插补，结果理想，平均判对率可达 85%。

关键词：旱涝等级；列联表；概率；插补。

一、引 言

对历史旱涝等级资料的分析有很多种．为了进行时间、空间上的规律分析，不仅采用一些计频方法，而且使用了许多将等级值作为量值的分析方法[1]．虽然从严格定义讲，旱涝等级不具有量值意义，而是状态的标号，但这种标号却相应根据旱涝程度有序地对应着量值序列[2]．这种有序数列用量值分析方法并非完全没有意义．但在计频方法中，将等级值作为状态标号处理才是比较严格的．

中国气候史料丰富，为研究过去气候规律提供了得天独厚的条件，但由于历史久远，各种因素的影响，存在许多地区上某些年代的缺佚．由此而来的旱涝等级序列存在相当数量的缺值．这种空间上和时间上的序列不连续，造成了分析历史气候规律的困难．以往的补缺方法，是根据旱涝等级的空间等值线读出相应缺值站的等级[3]．这种方法事实上也是建立在量值分析方法的基础上．因为做等值线图就是将等级值作为量值看待．由于一般地说，气候状况有一定的空间连续性，所以这种插值方法虽然不十分严格，却可以保证一定的准确率，并且是方便易行的．

本文应用的方法是将等级做为状态标号，因而在数学处理上比较严格，插值完全建立在空间旱涝等级的内部结构相关联规律上．实例计算时，考虑到资料的可靠性、连续性、代表性和空间分布，从 120 个站的旱涝等级资料[4]中选取 13 个站的资料进行分析（站点见表 1）．为了使资料更加可靠，作者用清代雨雪分寸奏折的资料(1736—1911 年)对旱涝等级值进行校正，故选用年代为 1736—1979．序列缺值共 219 个，占序列应有值的 6.9%．

1993 年 3 月 29 日收到，4 月 27 日收到修改稿．
* 国家自然科学基金委员会、中国科学院联合资助项目．
1）资料基本来自《中国五百年旱涝分布图集》的旱涝等级值，其中 1736—1911 年的值与清代雨雪分寸奏折的资料相互校正过，1911 年后完全用原值．

二、方法和原理

1. 方法步骤

设某台站K年旱涝等级序列为 $Y=y_1,\cdots,y_K$，其余 l 个台站 $X_1,\cdots,X_l$ 的K年旱涝等级序列分别为：

$$x_{1,1},\cdots,x_{1,K},$$
$$\vdots$$
$$x_{l,1},\cdots,x_{l,K},$$

所有台站旱涝等级全为5级，现欲用相关列联关系来估计Y台站某年的旱涝等级，步骤如下：

1）对 $(Y,X_1),\cdots,(Y,X_l),K$ 年数据分别建立5×5列联表[5,6]：

(Y,X_i) 5 × 5 列联表

		Y 1	2	3	4	5	$n_{i\cdot}^{(i)}$
X_i	1						
	2						
	3			$\cdots,n_{im}^{(i)},\cdots$			
	4						
	5						
	$n_{\cdot m}^{(i)}$						$N^{(i)}$

表中 $n_{im}^{(i)}$ 表示K年内 X_i 取 l 级，Y 取m级的频数；$n_{\cdot m}^{(i)}=\sum_{l=1}^{5}n_{lm}^{(i)}$；$n_{l\cdot}^{(i)}=\sum_{m=1}^{5}n_{lm}^{(i)}$；

$N^{(i)}=\sum_{m=1}^{5}n_{\cdot m}^{(i)}=\sum_{l=1}^{5}n_{l\cdot}^{(i)}=K;l,m=1,2,\cdots,5;i=1,\cdots,I$。

2）计算当 X_i 取 l 级时，Y 站取m级之条件概率：

$$\theta_{lm}^{(i)}=P(Y=m|X_i=l)=\frac{n_{lm}^{(i)}}{n_{l\cdot}^{(i)}}.$$

3）计算 $P_{lm}=\sum_{i=1}^{I}c_i\theta_{lm}^{(i)}$，

c_i 为条件概率的加权平均数，由下式计算：

$$c_i=\begin{cases}0, & \text{当 } E_i\leqslant e\\ N_{lm}^{-1}, & E_i>e\end{cases}$$

式中 e 为一预先给定的正数，$e>0$；$E_i=\dfrac{P(Y=m|X_i=l)-P(Y=m)}{P(Y=m)}$

$= \frac{\theta_{lm}^{(i)} - \theta_{.m}^{(i)}}{\theta_{.m}^{(i)}}$; $\theta_{.m}^{(i)} = P(Y=m) = \frac{n_{.m}^{(i)}}{N^{(i)}} = \frac{n_{.m}^{(i)}}{K}$; N_{lm} 为 $E_i > e$ 的项数 $(i=1,\cdots,I)$.

4) $\hat{y}=m_t$ 为 $\max(P_{lm})$ 的相应级别,其中 l 为 $x_1,\cdots,x_I$ 在 Y 站补缺年份的等级值.

2. 原理依据

以上方法的理论依据为:

1) 用 $\max(P_{lm})$ 相对应的级别来估计缺值 y,是以极大似然估计思想为依据,表征了某年当 $x_1,\cdots,x_I$ 取定级别时,y 取哪一级的可能性最大.

2) E_i 体现了 X_i 已取定某 l 级时对同年 Y 取 m 级的影响程度. E_i 值大到一定程度才能参加 P_{lm} 的加权计算. 因为 $E_i > e > 0$,即为

$$\frac{P(Y=m,X_i=l)}{P(Y=m)P(X_i=l)} > 1+e$$

此种情形的联合概率 $P(Y=m,X_i=l)$ 大于两个事件 $Y=m$ 与 $X_i=l$ 单独发生的概率乘积,因此 $X_i=l$ 对于 $Y=m$ 有较大影响,只让这种情形的 X_i 参加 P_{lm} 的加权计算;或者说,当条件概率 $P(Y=m,X_i=l)$ 减小趋向全概率 $P(Y=m)$ 时,X_i 取 l 与 Y 取 m 无相关性, $P(Y=m,X_i=l)$ 对 Y 取 m 级的可能性无贡献,不参加加权平均概率计算(当 $P(Y=m,X_i=l)$ 小于 $P(Y=m)$ 为反相关,减少 Y 取 m 级的可能性,暂不考虑).

可以看到,序列 $Y=y_1,\cdots,y_K$ 中缺值的年份可以不止一个,如果 I 个台站也有空值年,只计 $(Y,X_1),\cdots,(Y,X_I)$ 共有值年(此时 $N^{(i)} \neq K$),并且设定若 k 年时 X_i 为空值时 $c_i=0$,则方法仍可进行.所以对于一个等级时空场序列,所有单站序列都可以轮换当做被插补序列 $Y=y_1,\cdots,y_K$ 用上述方法计算一次,原序列的空缺值即可以用计算出的序列值对应插补. $N_{lm}=0$ 时, $P_{lm}=0$,被插补站 k 年与其余站无相关性,不能插补,原为空值仍为空值. 与实际相符概率可认为等于计算出的等级序列相对于原序列的判对

表 1 计算序列与原序列对比判对率表(%)

年段 A.D.	北京	太原	西安	洛阳	济南	徐州	苏州	安庆	吉安	江陵	岳阳	杭州	福州	年段平均
1736-1765	83	83	93	85	83	88	77	77	84	93	83	91	88	85
1766-1795	83	80	90	100	83	81	83	77	77	87	86	73	75	83
1796-1825	87	77	70	80	87	93	96	83	81	87	86	77	91	84
1826-1855	83	77	80	81	87	90	81	87	83	87	79	92	90	84
1856-1885	77	93	90	80	77	63	86	70	85	83	96	71	92	82
1886-1915	87	87	87	83	83	83	83	77	80	89	95	86	93	86
1916-1945	87	90	87	83	90	90	95	90	96	100	86	91	93	91
1946-1979	79	85	82	79	79	81	85	88	89	88	88	91	85	85
站平均	83	84	85	84	84	84	86	81	84	89	87	84	88	总平均85
1916-1955	82	82	85	73	77	82	93	88	85	100	77	87	77	84
1916-1965	72	74	68	78	74	76	77	78	84	96	71	82	70	77

率．例如，某站30年中有2年缺值，经插补后的序列有24年与原序列相同，判对率为24÷28，约等于86%，那么，原序列的2年缺值用插补序列中对应的值代替，与实际相符的可能性为86%．

三、实例计算结果

考虑到气候变化中的类型突变，一般选取K值在20至50年之间，ϵ取值宜使$N_{l_m} \neq 0$，且使判对率较高．实例计算中，取 $K=30$ 年进行分段(最后一时段为34年)，$\epsilon=0.3$．与实际序列值相比的判对率（判对率为插补后与原序列相同的次数除以原序列有记录的总次数)见表1，总平均判对率为85%．因此以计算值补入原序列的219个缺值与实际情况相合的可能概率为85%．

四、几点讨论

(1) 本方法要求被插补站序列与其余站共有值年不能太少，一般至少保证与一个站共有值年不少于样本时段的一半．如取30年为一时段，被插补的站应至少保证与其它站中的一个共有值年多于15年．如果所取时段内达不到要求，可适当延长时段使之满足。但时段取得过长，在一个时段内气候类型有突变发生，那么空间相关性降低，使得判对率降低．我们做一个实验，第7个时段从1916年起，增加时段为40年，判对率为84% 尚无明显变化，而延长到50年，判对率立即下降为77%，见上表．那么可以说延长的这10(1955—1965)年内，气候类型发生了突变．因此这种方法还可以作为检验气候类型的一种方法，我们今后将进一步研究．

(2) 我们选取的空间列联相关概率是单一等级间的列联相关概率，而不是选空间站的全级相关．这是因为考虑到气候变化本身的规律．例如北京大旱时，济南为旱的相关性可能很好，但北京为正常级时，济南5个级都可能发生．这种现象是可能普遍存在的．

五、小　结

通过研究和实践，本方法有如下特点：

(1) 数学处理方法上相对比较严格．

(2) 揭示等级空间内部相关、遥相关、等级(或值域)相关等规律，插补值可信度高．

(3) 比应用贝叶斯定理[7]的类似方法减少很大计算量．

(4) 寻找等级相关，可能比量值相关更加符合气候变化的现象和规律．

(5) 方法中确定分时段间隔的手段尚待进一步改进，应克服人为因素的影响．

参　考　文　献

[1] 张家诚、张先恭、许协江，1983，中国近五百年的旱涝，气象科学技术集刊 4，气象出版社．

[2] 张德二，1983，重建近五百年气候序列的方法及其可靠性，气象科学技术集刊4，气象出版社．

[3] 郑斯中、冯丽文，1985，我国冷的时期气候超常不稳定的历史证据，中国科学B辑，第 11 期，1038 p.

[4] 中央气象局气象科学研究院等,1982,中国五百年旱涝分布图集,地图出版社.
[5] 柯惠新、黄京华、沈浩,1992,调查研究中的统计分析方法,北京广播学院出版社。
[6] 黄明星、陈水校,1991,应用加权列联表分析法预测旱稻纹枯病发生量,中国农业气象, 12 卷,第 1 期.
[7] *Robert L. Winkler*, 1991, 贝叶斯推断,大气科学中的概率统计和决策, [美] *Allan H. Murphy, Richard W. Katz* 编,史国宁,周诗健等译,222—238,气象出版社.

A Method of Space Correlation Contingency Probability for Filling the Missing Data in Chinese Historic Dryness and Wetness Grades Series

Wang Qidong

(Commission for Integrated Survey of Natural Resources, Chinese Academy of Sciences, Beijing 100101)

Xiang Jingtian

(Institute of Applied Mathematics, Chinese Academy of Sciences)

Abstract

By the method of Space Correlation Contingency Probability (SCCP), the missing data in Chinese Historic Dryness and Wetness grades series of thirteen Stations from 1736 A.D. to 1979 A.D. can be filled (the missing data account for 6.3% of total data in the series). The result is satisfactory, for the average reliability can reach 85%.

Key words: Dryness and Wetness Grade; Contingency Table; Probability; Filling missing data.