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

历史气候资料的几种统计分析方法

潘一民 项静恬
(中国科学院应用数学研究所，北京 100080)

提 要

本文提供了历史气候资料分析的几种统计分析方法，其中的探索数据分析法、滚动检验法、时序建模法可用于判断时间序列的结构变化；而多层拟合法和加权列联表法则能对定量数据或等级数据的缺值实现有效的插补。以上方法在与中科院地理所协作的古气候资料分析研究中得到了有成效的应用。

关键词：探索数据分析；时间序列；建模；插值。

在与中科院地理所协作进行古气候资料研究的过程中，我们提供了几种判断序列结构变化和实现缺值插补预报的统计方法。经实际数据的分析表明，这些方法不仅简便易行，而且应用有效。

一、气候序列结构变化的判断和检验

设已收集了某地区N年的气候资料，为检验长时期气候状况有无显著变化，我们设计了对总序列实行分段"滚动"考察的检验方法。具体办法是将总序列 $\{x_i\}$，$i=1,2,\cdots,N$ 按长度 n 为一期，以 n_0 为滚动间隔分解成若干子序列，其中 $\{x_1,x_2,\cdots,x_n\}$ 为第一期，$\{x_{n_0+1},x_{n_0+2},\cdots,x_{n_0+n}\}$ 为第二期，…，$\{x_{(l-1)n_0+1},\cdots,x_{(l-1)n_0+n}\}$ 为第 l 期。一般取 $n_0 \ll n$。例如取 $N=1000$，$n=100$，$n_0=10$。然后，对按上述方法实现分期的滚动子序列选用如下几种办法分析考察，将有助于我们判断气候序列的结构是否发生了变化。

1. 探索数据分析法

探索数据分析法是70年代末期产生和兴起的一种统计分析方法。其不同于经典统计方法之处，是它无需对数据进行统计分布的先验性假定，直接通过数据运算和绘制图形来观察数据，并为进一步进行实证性数据分析(即统计分析)提供初步结论和依据。下面将提供的是 Tukey 箱线图法，可用于对长期序列的结构变化进行直观判断，具体做法分以下步骤实现：

(1) 确定对总序列进行分期的样本长度 n 及"滚动"考察的间隔长度 n_0；

1993 年 4 月 5 日收到，5 月 9 日收到修改稿。
* 国家自然科学基金资助项目。

(2) 对间隔为 n，长度为 n 的各期子序列计算下列统计量（将数据从小到大升秩排列）：最小值 I；1/4 分位数 H；中位数 M；3/4 分位数 Q；最大值 A；

(3) 对每期子序列画 Tukey 箱线图(图 1)；

(4) 以时间 t 为横轴，数据值 x，为纵轴，在各期起点时间的位置上画出箱线图并进行比较(图 2)，即可直观判断序列结构的变化。

Jukey 箱线图还有其它画法与功能，参见文献[1]。

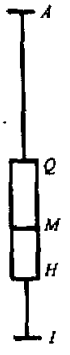

图 1 Tukey 箱线图

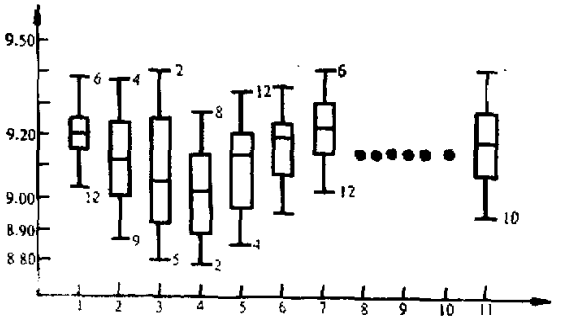

图 2 箱线图的滚动比较

上述利用 Jukey 箱线图分析数据结构变化的办法适用于连续变化型数据，下面给出一种对分段频数进行探索考察的办法，对于连续数据和等级数据都很适用，具体步骤如下：

(1) 统计出第 l 期子序列中第 j 级气候状况的数据个数 n_{lj}，并计算其频数 p_{lj}：

$$p_{lj} = n_{lj}/n,\ j = 1,2,\cdots,J;\ l = 1,2,\cdots,L.$$

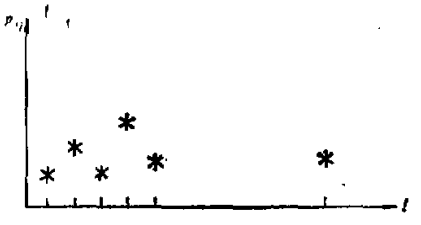

图 3 第 j 级频数分布图 ($j = 1,2,\cdots,J$)

(2) 以期号 l 为横轴，p_{lj} 值为纵轴作出 J 张同级气候在不同期的频数分布图（图 3），直接考察其起伏或周期变化情况。

此种方法在数据缺值时也能采用，只需在计算频数时分子分母减少相应个数即可。

利用探索分析所提供的初步结果，可对数据结构是否变异进一步作统计检验和时间序列分析。

2. 统计检验法

为了检验各期中气候状况有无显著差异，可采用经典的 Pearson χ^2 检验法。例如对于划分等级后的数据，若要检验滚动的第 l 期与第 m 期数据有无显著性差异，则构造统计量：

$$\chi^2=\sum_{j=1}^{J}\frac{(n_{mj}-n_m p_{lj})^2}{n_m p_{lj}}=\sum_{j=1}^{J}\frac{(n_l n_{mj}-n_m n_{lj})^2}{n_l n_m n_{lj}}$$

$$=n_m\sum_{j=1}^{J}\frac{(p_{mj}-p_{lj})^2}{p_{lj}},\quad n_m=\sum_{j=1}^{J}n_{mj}$$

根据中心极限定理和连续映射定理，当 n_l 和 n_m 足够大时，可以认为上述统计量 χ^2 接近于自由度为 $J-1$ 的 χ^2 分布．选取一定的置信水平 α（例如 $\alpha=0.05$），令 $P\{\chi^2_{J-1}\geqslant\chi^2_{J-1}(\alpha)\}=\alpha$，则当 $\chi^2>\chi^2_{J-1}(\alpha)$ 时，认为第 l 期与第 m 期的气候状况有显著差异，否则认为差异不显著．

如果希望集中考察某几个级别的差异，则只需把其余的级别统归于一个级别，然后仿上法进行统计检验即可．此时 χ^2 分布的自由度应为新级别个数减 1．

本方法是两个时期气候状况的静态比较法．

3. 时序建模分析法

为了动态地考察数据结构的变化，我们可通过时序建模来实现，具体步骤如下：

（1）确定对总序列进行分期的样本长度 n 和"滚动"考察的间隔长度 n_0；

（2）对各子序列建立时间序列模型[3]；

（3）比较相邻各期的模型结构，检验其差异是否显著．若阶数相同且系数相近，则认为差异不显著，否则通过 F 检验判别．F 统计量计算公式为

$$F=\frac{A_0-A_1}{S}\Big/\frac{A_1}{N-r}\sim F(S,N-r)$$

式中 A_0 和 A_1 为两个不同期数据序列相应模型的残差平方和（$A_0>A_1$），S 为两模型参数个数的差值，r 取高阶模型参数个数．

对预先给定的置信水平 α，由 F 分布表查出满足 $P(F\geqslant F_\alpha)=\alpha$ 的 F_α 值，若算得的 F 值大于 F_α，则认为两期序列结构存在明显差异．

对子序列的建模建议采用自回归 $AR(p)$ 模型，此法手段简单，可借用多元回归系数的估计程序，且参数个数即为模型阶数．

二、数据序列缺值的插补

气候资料常因这样那样的原因呈现时间或空间上的缺值，本文提供两种方法，能充分利用现有资料蕴含的信息，给出缺值的合理估计．

1. 时间序列缺值的多层拟合法

该方法按以下步骤来实现：

（1）计算序列中缺值的最大游程 l，以 l 为间隔提取原序列中子序列 $\{x'_t\}$：$x'_t=x_{t\times l}$ $(1\leqslant t\leqslant[N/l])$．

（2）对子序列 $\{x'_t\}$ 中所含有的原序列中缺值，用如下公式计算并代替：

$$x_i' = \frac{1}{2L - M + 1} \sum_{k=-L}^{L} x_{i+k}',$$

式中 S 为缺值在子序列中位置，M 为上式中和号内缺值个数，L 为求和限，其大小人为给定。

(3) 对序列 $\{x_i'\}$ 拟合时间序列模型。

(4) 用模型拟合值代替第 2 步中的 x_i'，并补回原序列中相应位置，作为该层补缺或修正。

(5) 若原序列缺值补齐则停止，否则计算该层补缺后剩余缺值最大游程，返回开头再进行下一层的计算和补缺。

取缺值最大游程作为提取子序列的间隔，可使每层补缺的个数尽可能少，以便让新补数值含有原序列中尽可能多的信息。且随着插补层次的增多，每层最大游程数单调递减，原序列中缺值也逐层减少，直至补齐，游程数 l 为零，迭代即可终止。

本方法适用于一维时间序列中存在较多缺值的情形，通过多层拟合，可逐步减少缺值个数，修正上一层的拟合数，并借助于时间序列模型，使缺值的插补趋于合理。

2. 应用加权列联表法实现插值和预报

加权列联表方法可以对多维等级数据中的缺值实现有效插补，也可用于某因素未来值的预报。

今设需插值或预报的气候资料为 Y，其余各种气候要素资料为 $X_1, X_2, \cdots, X_r$。前者称为预报量，是 L 等级数据；后者称为预报因子，分别具有等级 $L_1, L_2, \cdots, L_r$。为方便起见，预报量和预报因子均取同样长度的样本，记为 N。

加权列联表法的步骤如下：

(1) 对于 $(Y, X_1), (Y, X_2), \cdots, (Y, X_L)$ 分别建立 $L \times L_l$ 列联表，例如对于 L 和 L_l 等于 3，有 3 × 3 表(表 1)

表 1 (Y, X_l) 的 3×3 列联表

X_l \ Y (n_{ik})	1	2	3	$n_{i\cdot}$
1	2 (Q_{11}^l)	1 (Q_{12}^l)	0 (Q_{13}^l)	3
2	0 (Q_{21}^l)	6 (Q_{22}^l)	0 (Q_{23}^l)	6
3	0 (Q_{31}^l)	0 (Q_{32}^l)	4 (Q_{33}^l)	4
$n_{\cdot k}$	2	7	4	

(2) 用 χ^2 检验法判别 Y 与 X_l 相关性是否显著。

χ^2 值由下式计算：

$$\chi^2 = \left(\sum_{i=1}^{L_l} \sum_{k=1}^{L} \frac{n_{ik}^2}{n_{i\cdot} n_{\cdot k}} - 1 \right) N$$

式中N为样本长度，L为Y的等级数，L_I为X_I的等级数。n_{ik}为X_I取i级而Y取k级的频数，$n_{i\cdot}$为X_I取i级的样本数，$n_{\cdot k}$为Y取k级的样本数，均由表1可得。表1括号中Q值含义将于第4步介绍。

本检验自由度为$(L-1)\times(L_I-1)$，置信水平α取为0.05，若$X^2<X^2_{(\alpha)}$，则认为X_I与Y相关不显著，可以取消X_I对于Y的加权插值或预报。

(3) 求各因子X_I，$I=1,2,\cdots,r$的列联系数C_I：

$$C_I=\sqrt{X^2/(N+X^2)}$$

由此公式可见，C_I值大小与(Y,X_I)的相关性呈单增关系，列联系数C_I是对X^2值的合理修正。其作用在于消除样本含量的影响，提示变量间关系的真正密切程度[4]。C_I值将用于计算插值或预报值。

(4) 计算列联参数Q^I_{ik}，并记于表1括号内。

$$Q^I_{ik}=n_{ik}/n_{i\cdot}+n_{ik}/n_{\cdot k}$$

等式右边各值由表1提供。不难看出，Q^I_{ik}是两个条件概率之和，前者为$P(Y=k|X_I=i)$，后者为$P(X_I=i|Y=k)$，因此Q^I_{ik}是$Y=k$与$X_I=i$相关程度的一个度量，Q^I_{ik}越大，两者相关性越强。

(5) 计算加权列联参数和P_k，对Y的级别作内插或预报

$$P_k=\sum_{I=1}^{r}C_IQ^I_{ik},\quad k=1,2,\cdots,L.$$

式中和号内的Q^I_{ik}下标i为X_I已取定的级别。取L个P_k中最大值相应的等级为Y在某时刻上的预报或插补值，即

$$\hat{Y}=k_0,\quad P_{k_0}=\max_{1\le k\le L}P_k$$

若计算结果出现P_k值相近的情形，则可参照经验进行调整。

以上提供的方法与其它统计方法的配合应用，在古气候资料分析中发挥了有效作用。例如探索数据分析的箱线图对近五百年我国旱涝分区提供了一种划分手段；又如列联表法经改进后成功地用于历史气候资料时空缺值的插补，拟合率达85%；在用滚动时段的雨量频数进行主成分分析时，发现在1816、1833、1870年三处累计贡献率达85%的主分量数目有如下突然变化：1816年由4个减至2个，1833年由2个增至3个，1870年由3个增至5个。而用传统的Mann-Kendal方法分析数据，仅发现在1830年一处显示了突变，以上应用的详细分析将由地理所有关研究人员另文讨论。

参　考　文　献

[1] John W. Tukey, 1977, Exploratory Data Analysis, Addisonwesley Publishing Company.
[2] 复旦大学编，1980，概率论(第二册数理统计)，人民教育出版社.
[3] 项静恬等，1991，动态和静态数据处理——时间序列和数理统计分析，气象出版社.
[4] 柯惠新等，1992，调查研究中的统计分析方法，北京广播学院出版社.

Some Statistical Methods of Analyzing Historical Climatic Data

Pan Yimin Xiang Jingtian

(Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing 100080)

Abstract

In this paper, some statistical methods of analyzing long term climatic data are presented. Among them, exploratory data analysis, moving test and time series modelling may be used to decide the change of the structure of a time series; moreover, progressive fitting and weighted contingency table are effective on interpolation and for casting of quantitative data or qualitative data. These methods have been successfully applied to analysis of ancient climatic data in a research cooperative work with Institute of Geography, Chinese Academy of Sciences.

Key words: Exploratory data analysis; Time series; Modelling; Interpolation.