观测数据非线性时空分布理论和方法

封国林　董文杰　龚志强
侯　威　万仕全　支　蓉　著

气象出版社

观测数据非线性时空分布理论和方法

封国林　董文杰　龚志强
侯　威　万仕全　支　蓉　著

气象出版社

内容简介

本书针对气候系统的观测数据具有非线性、多层次结构和非平稳性的特征，系统地介绍了动力学相关因子指数（Q指数）、启发式分割算法（BG）、经验模态分解（EMD）、复杂度和幂律尾指数等一系列处理观测数据的非线性时空分布新的理论和方法，及其最新的相关研究成果，旨在为气候突变的检测和归因、观测数据信息的提取与预测以及从某种程度上区分自然变率和人为变率对20世纪增暖的贡献提供新的有效途径。

本书为从事与气候诊断和气候变化有关的研究工作者和研究生提供新的研究思路和分析方法；同时，也可供从事水文、地震、经济等其他非气象领域的科研工作者以及高等院校师生查阅参考。

图书在版编目(CIP)数据

观测数据非线性时空分布理论和方法/封国林等著．北京：气象出版社，2006.12

ISBN 7-5029-4234-3

Ⅰ．观… Ⅱ．封… Ⅲ．地理-数据-数理统计 Ⅳ．①P20②0212

中国版本图书馆CIP数据核字(2006)第142742号

出版者：气象出版社　　**地　址**：北京市海淀区中关村南大街46号

网　址：http：//cmp.cma.gov.cn　　**邮　编**：100081

E-mail：qxcbs@263.net　　**电　话**：总编室：010-68407112　发行部：010-62175925

责任编辑：李太宇　陆同文　　**终　审**：汪勤模

封面设计：张建永

印刷者：北京中新伟业印刷有限公司

发行者：气象出版社发行　全国各地新华书店经销

开　本：787×1192　1/16　　**印　张**：14.75　　**字　数**：372千字

版　次：2006年12月第1版　2006年12月第1次印刷

印　数：1～2000

定　价：40.00元

前　言

本书是根据作者近年来在非线性时空分布理论和方法方面的主要研究成果编撰而成的一本专著。

自20世纪70年代以来，非线性分析方法在大气科学中的应用越来越深入，对观测数据非线性时空分布特征的研究作为气候诊断与分析的一个重要研究方向，越来越受到气象科学工作者的重视。气候系统是一个开放、非平衡、高度复杂的动力学系统。它具有非线性、非平稳性和多层次性，在时间和空间上存在各种尺度相互作用和反馈的复杂性特征。并且不同时空尺度的现象具有彼此不同的特点。对于某种时空尺度现象，如果不以这种共同的认识为原则，找出其特殊的本质，就很难甚至无法成功地揭示、解释和预测这一现象。

作者通过多年来的教学和研究，并借鉴国内外在处理非线性时间序列方面新的研究途径，在中国气象局培训中心丑纪范院士、国家海洋局巢纪平院士和中国科学院大气物理研究所杨培才研究员等的鼓励和指导下，在同行和有关专家的帮助和支持下，终于完成了《观测数据非线性时空分布理论和方法》一书。

本书以尽可能精练的语言和公式，说明非线性时空分布理论中若干概念、理论和方法，并尽量做到理论与实际应用紧密结合，供读者在实际工作中由浅入深地理解和应用。

全书以气候突变和极端事件的检测与归因为主线，分别从统计和动力学结构两个角度进行研究。全书分为七章，第一章介绍时间序列的研究进展；第二章到第五章系统地介绍了新的非线性分析方法，如动力学相关因子指数(Q指数)、启发式分割算法(BG算法)、复杂度和幂律尾指数等，并对这些方法做了有效性检验，进而初步应用于实际气候资料的分析和突变检测等。目前，国内对于这些方法及其应用的研究还相对较少。第六章和第七章主要介绍了经验模态分解(EMD)等新的观测数据信息的提取与研究的方法，并在此基础上结合概率密度和非线性去噪等方法讨论了观测数据的均一性及动态建模等。

全书自成系统，章节之间既相对独立，又相互联系，同时还兼顾了在大气科学中的应用。为方便相关科研工作者对这一研究领域新的理论和方法有一个较全面和深刻的理解，在编写过程中，作者既注意了全书的可读性，力图以简洁的数理描述，清晰的图像，结合理想序列试验，使各种方法的介绍通俗易懂；同时从解决某个气候研究中的热点问题出发，将各种方法应用于观测数据或代用资料的检测，并结合前人的工作对结论进行讨论和分析，进而保证了内容的科学性；本书中所涉及的各种非线性分析方法均具有清晰的统计和物理意义，因此具有

广泛的适用性和实用性，不仅适用于大气科学领域的研究，同时也适用于地理学、物理海洋学、经济学以及信号处理等研究领域。

总之，非线性时空分布理论和方法与实际应用研究的有机结合，是本书的突出特点。书中还针对气候突变和极端事件的检测与归因做了深入的探讨。本书尝试为非线性时空分布理论和方法的研究向实际应用的转化提供范例，同时希望作者与读者相互交流，相互启发，促进在本研究领域中认识的进一步深化。

本书在国家重点基础研究发展计划“北方干旱化与人类适应(2006CB400500)”、国家自然科学基金重大研究计划“全球变化与区域响应(90411008)”项目和扬州大学出版基金的共同资助下完成。在完成本书的过程中，得到了高新全博士、何文平博士、张文博士和邹明玮博士的热情支持和帮助，他们不仅为本书提供了不少有价值的研究成果，而且花费了大量的精力进行审校，在此致以深深的谢意！

在本书的出版过程中，得到了气象出版社李太宇和陆同文等同志的大力帮助，特此感谢。鉴于本书涉及领域较广，书中难免有欠妥之处，敬请读者和广大科技工作者批评指正。

作者

2006年7月

目 录

1 时间序列的研究进展

一般来说，时间序列是指在离散参数(时刻) $t_1 < t_2 < \cdots < t_i \cdots < t_n$ 上得到的以 t 为自变量(离散参数) 的有序数值的集合，记为

$$x(t_1), x(t_2), \cdots, x(t_i), \cdots, x(t_n) \tag{1.1}$$

有时简记为 $\{x_i\}$ 或 $x_1, x_2, \cdots, x_i, \cdots, x_n$。例如，在气象记录中，采用自动化或自记仪器在时间坐标上连续读取或记录的气象要素变化曲线，虽然在连续时间域内都有记录，但在研究分析时通常以固定的时间间隔加以采样，使其成为离散化的数字记录。所有这些在时间上有序的气象记录，都可认为是气象随机过程的观测结果，即样本函数。它们的共同特点是，以离散时间为参数的数据记录序列或数字记录序列，通常把这种记录序列叫做时间序列。

在实际工作中，常常不需要了解或不可能得到连续参数过程的全部可能记录，而只能获得离散参数过程的观测记录或连续参数过程的离散化采样记录。因此，在实际工作中广泛应用离散参数随机序列，对它们所使用的许多行之有效的统计方法，其理论基础虽来源于随机过程，但在数学处理上却常常可以避免某些复杂性。因而在自然科学和社会科学的许多领域，实际上都是以“时间序列”作为研究对象。[1] 近几年非线性科学的飞速发展极大地推动了时间序列的研究，本章将扼要地介绍时间序列的研究进展，并着重介绍非线性领域中有关突变检测与归因的数理方法。

1.1 动态数据预处理[2]

由于数据存在观测误差、仪器误差和随机误差，例如，由于数据传输系统中发生信号失真或丢失等而产生的噪声数据，通常会在以后的时间序列分析中带来额外的误差，影响所建立的时间序列模型的精度。因此，在建立模型之前，必须先对动态数据进行必要的预处理，即质量控制，以便剔除那些不符合统计规律的异常样本，并对剔除异常样本后数据的统计特性进行检验，以确保时间序列模型的可靠性和置信度，并满足一定的精度要求。

关于数据预处理的具体方法和理论已经非常成熟，本节只扼要地介绍动态数据的预处理，主要包括平稳性检验、正态性检验、周期性检验和独立性检验。此外，还要对某些确知规律性的数据进行趋势性检验，以便保持被统计数据的纯随机性质。

一般来说，传统分析方法或近几年出现的一些“非线性方法”，都是假设观测资料是平稳的，因此，平稳性是传统时间序列分析、建模的基础和重要前提。时间序列的平稳性包含两个基本的假设：(1)序列的均值和方差为常数；(2)序列的自相关函数仅与时间间隔有关，而与此间隔端点的位置无关。在数理统计分析过程中，尤其是气象观测资料的平稳性检验中，通常应用游程检验法(或轮次检验法)，这是一个非参数检验法，其突出的优点在于它仅仅涉及一组实测数据本身，而与数据的分布类型无关，具有很好的实用性。

正态性是动态随机数据最重要的统计特性，目前常用的时间序列模型就是建立于具有正态分布特性的白噪声基础上。

正态分布的概率密度函数(PDF)可记为：

$$P(x) = (2\pi\delta^2)^{-\frac{1}{2}}\exp\mid(x-\mu)^2/(2\delta^2)\mid \tag{1.2}$$

其中 μ,δ^2 分别为样本总体的均值和方差。概率分布是概率密度函数的积分

$$\begin{aligned} P(x<X) &= (2\pi\delta^2)^{-\frac{1}{2}}\int_{-\infty}^{X}\exp[-(x-\mu)^2/(2\delta^2)]\mathrm{d}x \\ &= (2\pi)^{\frac{1}{2}}\int_{-\infty}^{(X-\mu)/\delta}\exp\left(-\frac{1}{2}x^2\right)\mathrm{d}x = \Phi[(X-\mu)/\delta] \end{aligned} \tag{1.3}$$

随机变量处于 α,β 之间的概率为：

$$P(\alpha\leqslant x\leqslant\beta) = \Phi[(\beta-\mu)/\delta]-\Phi[(\alpha-\mu)/\delta] \tag{1.4}$$

“卡埃平方(χ^2)拟合优度检验”是一种检验动态数据正态性的有效方法。它是利用 χ^2 统计量作为观察到的PDF和理论PDF之间偏差的量度。通过分析 χ^2 的样本分布可以用来检验两者是否相同。

从理论上讲，只要样本数据总量足够多，直至无限长，并符合正态性的随机序列一定具有统计独立性。例如，均值为零、方差为 δ^2 的正态序列，它的自相关系数一定满足 $\rho(\tau)=\delta(\tau)$，从而，说明样本数据之间具有独立性。

事实上，对于气象观测资料尤其器测资料，显然无法满足上述样本数据总量足够大的要求，因此，对于时间序列的样本数据来说，除了要检验其平稳性和正态性之外，还应该检验其独立性。

对于正态独立分布的随机变量，其自相关系数

$$\rho(r) = \delta(r) = \begin{cases} 1, & r=0 \\ 0, & r\neq 0 \end{cases} \tag{1.5}$$

因此，当 $r\geqslant 1$ 时，$\rho(r)=0$。但我们所能得到的是估计的样本自相关系数 $\rho(r)$，它一般不会等于 $\delta(r)$。如何从估计的自相关系数判断“真正”的函数是否满足独立性条件，可以利用Bartlett公式。该公式指出：若 $\rho(r)$ 在 $r>M$ 时趋于零，则当 N 足够大时其方差为

$$Var[\rho(r)] \approx \frac{1}{N}\sum_{m-M}^{M}\rho^2(m),\ r>M \tag{1.6}$$

而且，$\rho(r)$ 当 $r>M$ 时，近似于正态分布。

在有些情况下可能由于偶然因素，个别 $\rho(r)$ 在 $r>0$ 时，超出上式所约束的范围，有人提出另一种检验 x 是否独立的整体检验方法。构造统计量为

$$Q = N\sum_{r=1}^{k}\rho^2(r) \tag{1.7}$$

以“$\{x_i\}$ 为白噪声”做原假设，以 α 为显著性水平，则根据 α 和自由度 k，由 χ^2 分布表可查出相应的 $\chi^2_\alpha(k)$ 值，并与计算出的 Q 值比较。当

$$Q\leqslant\chi^2_\alpha(k) \tag{1.8}$$

时，则肯定原假设，即在 $(1-\alpha)$ 的置信水平上接受 $\{x_i\}$ 为独立的假定。若

$$Q>\chi^2_\alpha(k) \tag{1.9}$$

则否定原假设。式(1.8)即是检验判别式。

如果在时间序列中存在周期性或准周期性样本数据，则它们反映到时间序列的功率谱 $S(\omega)$ 上就会出现尖峰，因此，周期性或准周期性样本数据的功率谱很容易与随机数据的功率

谱(如钟形)相区分。

另外,也可以用自相关系数 $\rho(r)$ 进行周期性检验,此时,周期性数据序列的自相关系数呈连续振荡波形,而随机性数据的自相关系数则表现为单调的下降曲线。

如果能获得时间序列样本数据 PDF 的直方图,则可以根据其不同的形状来分辨其是否具有周期性或随机性。因为周期性或准周期性数据 PDF 的直方图呈下凹形(盆形),而随机数据 PDF 直方图却呈上凸形(钟形)。但当周期信号的方差比随机部分的方差小,或者包含一个以上的周期信号时,就不容易在直方图上判别。

时间序列分析都是假定样本数据来自平稳的和各态遍历的随机过程,也就是它们的期望值(均值、方差、相关系数等)都不随时间推移而变化,而且可以用时间平均代替总体平均,即分布的总体是不变的。当然,这样处理的前提条件是应当有足够长的数据记录才能使样本平均有代表意义。当任何一种平稳性条件被破坏时就出现非平稳。在实际情况中经常出现的有三种非平稳过程:均值非平稳、方差非平稳以及均值方差非平稳。

非平稳趋势检验对于单调的趋势是有效的,但在有些情况下也存在一定的局限性。例如,序列方差变化在正跳部分有很多逆序,但在随后的负跳部分逆序很少,而整个数据的逆序检验却可以是正常的;周期性趋势也可以看做一种特殊的非平稳趋势,因此,有时需要在某一时间序列中去掉一个线性的或缓慢变化的趋势,这种趋势项可能是由于数据中的某些分量是通过积分产生的。积分可以导致两种误差,首先是如果零点没有调准,则在每一采样时刻都有一个小的误差项,经过积分后这一常数项就变成了直线。这一线性趋势在谱分析或其他计算中会导致很大的误差。

另一类误差的产生是由于积分或低频噪声引起功率放大作用,而在数据中常有这类噪声,其经过积分后变成缓慢变化的随机信号,其变化速度在某种程度上取决于采样间隔。趋势项也并非都是误差,它可能代表时间序列中包含的有用信息(例如,前面提到周期性趋势),由于它的出现使过程具有非平稳性,因此,在对数据作平稳化预处理时需要提取出趋势项。一般来说,变化着的趋势项可以用滤波器来消除,而多项式形式的趋势项可以用最小二乘法来提取。

对于奇异数据的剔除目前还没有找到很好的方法,在一般情况下都是依靠分析人员的实际经验,再通过人工剔除的方法来进行。当然由于人为因素不可避免,往往会影响建模的精度,带来模型的系统误差。[3]

下面介绍一种线性外推的方法来对奇异数据进行剔除。该方法采用两个数字低通滤波器,它的输出是对输入函数的平滑估计,这里认为正常的数据是"平滑"的,而奇异点是"突变"的。样本方差的更新值为

$$\sigma^2(i) = \overline{x^2(i)} - \overline{[x(i)]}^2 \tag{1.10}$$

其中 $\overline{[x(i)]}^2$ 是先对数据作平滑处理后再平方得到的值,$\overline{x^2(i)}$ 是先对数据取平方后再作平滑处理得到的值。取 $\sigma^2(i)$ 的开方,可得标准差 $\sigma(i)$。接着是检查下一个数据点 $x(i+1)$,如果

$$\overline{x(i)} - ks(i) < x(i+1) < x(i) + ks(i) \tag{1.11}$$

则认为 $x(i+1)$ 是可接受的。k 是根据情况设定的适当数值,通常为 3 ~ 9,开始时不妨取为 6。如果 $x(i+1)$ 被认为是异点。则可用 $\hat{x}(i+1)$ 来替代,即

$$\hat{x}(i+1) = 2x(i) - x(i-1) \tag{1.12}$$

这实际上是线性外推。

这种方法必须附加一些参数设定，即事先规定连续外推的次数，以免出现无休止的外推。因为接连检测到一些异点后，最终的外推结果可能偏离很远，以致会排除本来正常的数据点。

1.2 时间序列模型分析方法[4]

分析和处理时间序列是为了从中提取有关的信息，揭示时间序列本身的结构与规律，从而认识产生时间序列的系统的固有特性，掌握数据内部系统与外部的联系规律，了解过去，预测未来。[5]

基于观测资料是平稳的假设，常见的分析时间序列的模型有自回归模型(AR 模型)，滑动平均模型(MA(q)模型)，趋势性和季节性模型(ARIMA)等。而常见的气候统计预测模型可以概括为以下几大类：

(1)时间序列模型。它是描述序列自身演变规律的模型，一般包含趋势项、周期项和随机项三部分。随机项通常用线性模型来描述，这类模型包括自回归(AR)、滑动平均(MA)、自回归滑动平均(ARMA)、自回归求和滑动平均(ARIMA)模型等等。其中发展较为完善的模型是 ARMA(p,q)，可表达为：

$$x_t = \sum_{i=1}^{p} \varphi_i x_{t-i} + a_t - \sum_{i=1}^{q} \theta_i a_{t-i} \tag{1.13}$$

其中 p,q 分别为 AR 模型和 MA 模型的阶数。若 $\theta_i \equiv 0$，式(1.13)变为 AR 模型；若 $\varphi_i \equiv 0$，则(1.13)式变为 MA 模型，它包含了 AR 模型和 MA 模型的特性。

在实际预测工作中逐渐发现数理方法尽管十分完善，但在天气预报和气候预测中有时也会遇到巨大困难，即所谓的“预测瓶颈”。20 世纪 70 年代，随着非线性科学的发展，科学家们尝试利用非线性科学的新成果对一些传统的方法加以发展和改善，逐步形成了门限自回归(TAR)、马尔柯夫链等非线性模型。实质上，上述方法还是基于时间序列的平稳性假设。这些模型在许多气象统计预测专著中均有较详细的介绍，本书不再赘述。魏凤英[4]提出了用多元分析手段解决时间序列预测问题的均生函数模型，为多步短期气候预测开辟了一条新途径。

(2)动态系统模型。气候系统作为一个随机系统，它的状态大多并不是严格平稳的，本质上是非平稳的，因此，Kalman 滤波可以用于描述非平稳的系统，它实质上是用一个最优化的递推数据处理算法建立自适应模型。Kalman 滤波目前多用于短期天气的 MOS 预报中，也有人尝试用在短期气候预测中。

另外，动态系统的多层递阶预测模型亦在气候预测中广泛使用。它的基本思想是把具有时变参数的动态系统的状态预测，分离成对时变参数和系统状态这两部分的预测，克服了回归方法中用固定参数模型来预测动态系统状态的局限性。

(3)多元回归模型。在气候预测中应用十分广泛的多元回归模型是在系统的动态方程不清楚的情况下，描述变量之间线性关系最有效的数学模型。其一般表达式为：

$$y = b_0 + \sum_{k=1}^{m} b_k x_k \tag{1.14}$$

其中 y 为因变量(预报量、预报对象),x_k 为影响 y 的自变量(预报因子)($k=1,\cdots,m$),b_0 为回归常数,b_k 为回归系数。通常采用最小二乘法来估计回归系数。

选择最优回归方程一般采用逐步回归的方法,不过,在计算机资源十分丰富的今天,完全可以从所有可能的子集回归中选择最优回归。针对不同预测问题的要求和数据存在的缺陷,发展出了与最小二乘法估计不同思路的其他回归方法,例如主成分回归、特征根回归及岭回归等模型。

(4)变量场预测的方法。气候预测中经常遇到变量场水平分布的预测问题。预测对象是一个空间变量场,因子也为空间场。以单点资料为基础的回归分析,局限于单点气候变量变化的统计规律,没有考虑点与点之间的相互联系,导致水平分布预测结果有时出现无法解释的跳跃。因此,变量场水平分布预测可以采用变量场展开的统计方法。其思路是把变量场展开成各种典型的特征向量与其时间系数的乘积和。假定在一定时间内,空间典型向量是稳定不变的,这时典型特征向量的系数变化反映了变量场随时间的变化。只要预测出未来时刻的时间系数,乘以典型特征向量就可以得到变量场的预测。常用的变量场预测展开方法有:经验正交函数展开(EOF)、车贝雪夫多项式展开及典型相关等。近年来王革丽等[6]提出了场时间序列预测的新思想和新理论,并在此基础上给出了时间序列试验分析,可以提高单点时间序列的"遍历性",从而提高预测精度。

(5)神经网络。近几年国内外文献中出现了将神经网络用于气候预测的研究成果。神经网络方法是目前国际上的热点学科之一,其包含的内容十分广泛,算法也十分繁多。它以其独特的结构和处理信息的方式,在许多应用领域都取得了显著的成效,特别是在处理非线性问题上显示出较强的能力。神经网络预测模型的参数是网络对原始数据进行不断学习、训练得到的。神经网络技术是人工智能的一个分支,虽然其预测模型也是用历史观测数据来构建的,但它并不属于概率论与数理统计的范畴,本书将不涉及这部分内容。

1.3 气候突变

在 1.2 节中扼要地介绍了一些传统的时间序列分析方法,在实际的应用过程中,假如数据样本足够大,还有可能出现下述情况:序列的长期变化常常表现为升降趋势交替变动,而并非长久的线性趋势,在两种明显不同的大趋势之间存在转折点。另一方面,序列的非平稳性还时常表现为数据取值状态的突然变化,即所谓突变。例如,平均值在很短的时间内从一种数值变为另一种数值。这种现象在气候时间序列资料中时常会遇到,需要用某些定量的方法加以识别并提取它的信号特征。本节主要介绍时间序列的转折和突变。

近代气候学研究把气候看做是不断变化的,这与经典气候研究有着本质的区别,导致了新气候概念的产生,新研究方法的提出,新研究事实的发现,新认识和思想的诞生。20 世纪 60 年代中期以来,以 R. Thom 的工作为先导而逐步建立的突变理论,目前,已被广泛应用于气候、地震等各个研究领域。[7~8]从历史资料来看,全球气候已经经历了各种时间尺度的巨大变化。可以预料,未来还将变化不息。气候不是一成不变的,它的变化具有阶段性。气候从一个阶段到另一个阶段的变化有两种基本形式,渐变和突变。气候的渐变表现在相当一段时间内在某一相对稳定态附近的振动,气候处于同一性质中;突变则是相对稳定态的不连

续跳跃，气候性质发生了根本改变。气候突变(Abrupt Climate Change)又称气候变化的不连续性、气候的跳跃，是普遍存在于气候系统中的一个重要现象。[9~10] 自从 Lorenz 和 Charney从理论上揭示了气候突变的可能性后，有关气候突变的研究得到了广泛的开展。一般认为，气候系统内部动力结构发生演化或外界的扰动过大，将导致系统的状态在相空间中不再趋向于原来的吸引子而是趋向于新的吸引子即发生了突变，它是一种多时间尺度的现象，可以发生在季节、年际、年代际、百年际甚至更长的时间尺度上，是气候系统所具有的非线性特殊的表现形式。[11~12]气候突变的研究对认识气候变化的性质及其气候预测有重要意义。所以，气候突变作为非线性变化的一个重要规律，越来越受到人们的重视。气候突变的研究主要从统计和动力学两个角度入手，目前许多研究仅仅着重于前者，对后者的研究国内外刚刚起步。

1.3.1 突变的统计学表征

(1)均值突变

在气候突变研究中，如果考察气候要素的统计量ξ是均值，则这种突变定义为均值突变。在样本 $x_1, x_2, \cdots, x_n$ 独立且服从方差 δ^2 正态分布的前提下，原假设 $H_0: Ex_1 = Ex_2 = \cdots = Ex_n$，假设 H_1：对于某个 m，$1 \leqslant m \leqslant n$，及 $\alpha_1 \neq \alpha_2$ 有 $Ex_1 = Ex_2 = \cdots = Ex_{m-1} = \alpha_1$，$Ex_m = \cdots = Ex_{n1} = \alpha_2$，其中 m 未知，信度 α 为给定值，如果检验肯定了 H_1，则 m 为突变点(见图 1.1)。[13]

(2)方差突变

如果考察的气候要素的统计量 ξ 是气候变率，则这种突变称为方差突变(见图 1.2)。

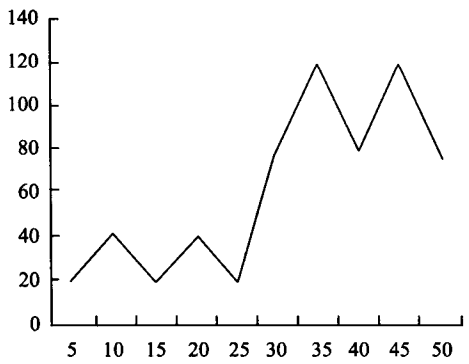

图 1.1 均值突变示意图[14]

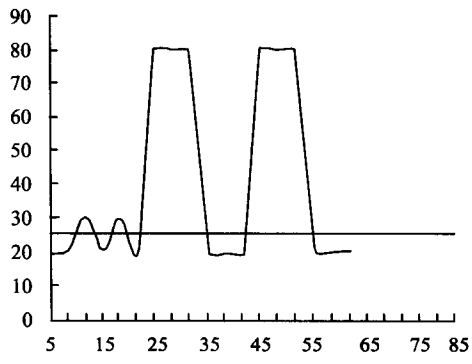

图 1.2 方差突变示意图[14]

(3)趋势突变

温度、降水等气候要素在某一时段持续减少(增加)，然后突然在某点开始持续增加(减少)的现象成为趋势突变(见图 1.3)。[13]

(4)频率突变

在气候系统中，同一气候事件由于受到不同的外强迫或者气候系统本身动力学结构发生变化，其出现的周期可能会相应地发生变化，对应着不同频率的变化，即频率突变。频率

突变是更为常见的一种气候突变形式(见图 1.4)。至于趋势突变是渐变中的突变形式,大的冰期与间冰期之间的变化通常表现为这种形式。[13]

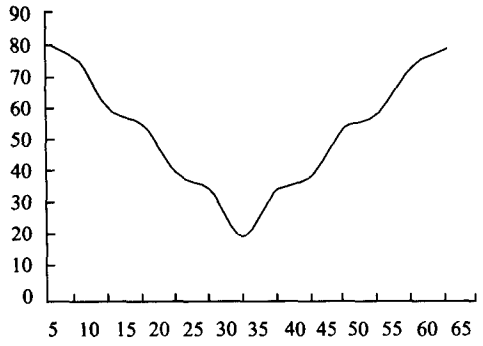

图 1.3 趋势突变示意图[14]

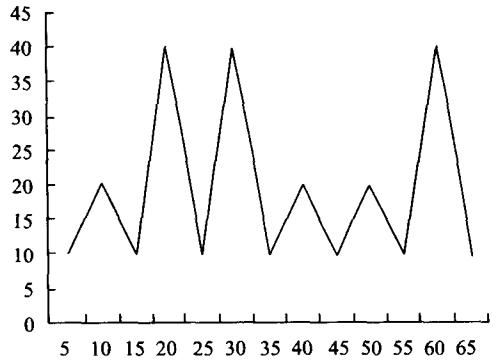

图 1.4 频率突变示意图[14]

(5)回归突变

在气候突变中,如果考察的气候要素的统计量 ξ 是回归系数,则这种突变即为均值回归突变。有自变量 $x_1, x_2, \cdots, x_n$ 和因变量 y,它们服从线性回归方程,但回归系数在 m 处有一次突变,即:

$$y(i) = \begin{cases} a_0 + a_1 x_1(i) + \cdots + a_p x_p(i) + \varepsilon_i, & 1 \leqslant i \leqslant m \\ b_0 + b_1 x_1(i) + \cdots + b_p x_p(i) + \varepsilon_i, & m \leqslant i \leqslant n \end{cases} \tag{1.15}$$

若两个向量 $A = (a_0, a_1, \cdots, a_p)^T, B = (b_0, b_1, \cdots, b_p)^T$ 不相等,$(\quad)^T$ 表示矩阵转置,则 m 为一个回归突变点。[13]

(6)概率突变

在气候突变中,如果考察的气候统计量 ξ 是事件发生的概率,即某一事件发生的概率在某一时刻有了突出的变化,即为概率突变。

(7)分布突变

在气候突变中,如果考察的是气象要素观测值同一时刻的分布,在某一时刻分布有了突然的变化,则称为分布突变。比如在一定时期,全球观测站点的某要素的同一时刻观测值遵循某种分布,而在某一时刻突变成另一种分布,即发生了分布突变。均值、方差等突变模型中与此不同的是观测值的分布在突变点前后都保持不变。[13]

1.3.2 传统的突变检测方法

目前在气候变化中,检测突变的方法主要有:(1) 滑动 t 检测法;(2) Gramer 法;(3) Yamamoto 法;(4) Mann-Kendall 法以及 Fisher 最优分割法和最大概率突变点检验方法等。下面对几种常见的突变检测方法的主要思想作简单介绍。[15~17]

(1)滑动 t 检测法

滑动 t 检测是通过考察两组样本平均值的差异是否显著来检验突变。其基本思想是把一气候序列中两段子序列均值有无显著差异看为来自两个总体均值有无显著差异的问题来

检验。如果两段子序列的均值差异超过了一定的显著性水平，可以认为均值发生了质变，即突变发生，在数学物理的角度，表现为相位的不同。

这一方法的缺点是子序列时段的选择带有人为性。为避免任意选择子序列长度造成突变点的漂移，具体使用这一方法时，可反复变动子序列长度进行试验比较，提高计算结果的可靠性。[13]

(2) Gramer 法

Gramer 方法的原理与 t 检验类似，区别仅在于它是用比较一个子序列与总序列的平均值的显著差异来检测突变。由于这一方法也有人为地确定子序列长度等因素，因此，在具体使用时，应采取反复变动子序列长度的办法来提高计算结果的合理性。[13]

(3)Yamamoto 法

Yamamoto 方法是从气候信息与气候噪声两部分来讨论突变问题的。Yamamoto 最先将信噪比用于确定日本地面气温、降水、日照时数等序列的突变，故称其为 Yamamoto 方法。Yamamoto 方法也是用检验两子序列均值的差异是否显著来判别突变的。从形式上它比 t 检验更简单明了。但它也存在与 t 检验相同的缺点。由于人为设置基准点，子序列长度的不同可能引起突变点的漂移。因此，应该通过反复变动子序列的长度进行试验比较，以便得到可靠的结论。[13]

(4) Mann-Kendall 法

Mann-Kendall 法是一种非参数统计检验方法。前面的三种统计检验方法都是参数方法，即假定了随机变量的分布。非参数检验方法亦称无分布检验，其优点是不需要样本遵从一定的分布，也不受少数异常值的干扰，更适用于类型变量和顺序变量，计算也比较简便。由于最初由 Mann 和 Kendall 提出了原理并发展了这一方法，故称其为 Mann-Kendall 法。但是，当时这一方法仅用于检测序列的变化趋势。后来经其他人进一步完善和改进，才形成目前的计算格式。这一方法的优点在于不仅计算简便，而且可以明确突变开始的时间，并指出突变区域，因此，是一种常用的突变检测方法。[13]

1.4 非线性时间序列的分析[18]

20 世纪 70 年代末，以混沌理论为核心的当代非线性科学得到了迅速发展，有力地推动了时间序列的分析研究。[18]人们发现，即使是一个十分简单的、完全确定的非线性系统，在一定的条件下也可以表现出非常复杂、非常随机的性质。动力学意义上的非线性时间序列分析开创于 20 世纪 80 年代初，它以重建相空间为基础，研究相空间动力轨道的性质，并据此进行预测。这类方法在本质上是动力学的、非线性的，在观念和方法上都有原始性的创新，非线性科学的思想和手段迅速地被应用到时间序列分析领域，形成了非线性时间序列分析这一新型的学科分支。[19~20]简单地说，凡是不能表示成式(1.16)都称为非线性时间序列。

$$x_i = \sum_{j=0}^{\infty} \beta_j \varepsilon_{i-j} \tag{1.16}$$

其中，系数序列$\{\beta_j\}$，满足

$$\sum_{j=0}^{\infty} \beta_j^2 < \infty \tag{1.17}$$

而$\{\varepsilon_i\}$是白噪声序列，满足

$$E(\varepsilon_i)=0,\ E(\varepsilon_i^2)=\delta^2<\infty \tag{1.18}$$

但这样定义非线性时间序列缺乏理论上的严谨。实际上，在非线性时间序列的分析中，我们主要研究平稳非线性序列，自20世纪70年代末以来，发展的非线性预测方法均是基于这样一种假设。[21]

1.4.1 传统方法

对于非线性时间序列的研究，大致可分为两种方法。一种方法主要是研究一般平稳非线性序列。它又有时域方法和频域方法之分。例如，高阶矩描述方法、高阶谱描述方法、Volterra展开方法等。[21]

高阶矩描述方法是一种时域方法。它通过引进序列的高阶矩来刻画其结构特征。例如，引入三阶矩函数连同序列的自协方差函数来描述序列一至三阶矩的相依结构。类似地，可以考虑比三阶矩更高的矩函数(例如第7.4节)。在实际中，一般只考虑到有限阶的矩函数。因此，它们都只是序列部分的结构信息。而且由于计算的困难，也不能考虑阶数太高的矩函数。高阶谱描述方法与高阶矩描述方法等价，是一种频域方法，并借助于高阶矩的谱来表示的。[22]

Volterra展开方法是将平稳序列$\{x_i\}$展开成如下形式的求和项：

$$x_i=a+\sum_{j=0}^{\infty}a_j\varepsilon_{i-j}+\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}a_{ij}\varepsilon_{i-t}\varepsilon_{i-j}+\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}a_{ij}\varepsilon_{i-t}\,\varepsilon_{i-j}\,\varepsilon_{i-k} \tag{1.19}$$

式中$\{\varepsilon_i\}$为白噪声序列。它也属于时域方法。显然，式(1.19)是线性序列式(1.16)的直接推广，但是，无论是在理论研究方面还是在应用方面，都只能考虑式(1.19)中的有限项求和，以此作为$\{x_i\}$的近似展开。与前述的高阶谱和高阶矩方法相似，这种近似展开也只能描述序列的部分结构信息。而且，即使只考虑到三阶项$a_{ij}\varepsilon_{i-t}\,\varepsilon_{i-j}\,\varepsilon_{i-k}$的求和项，在实际应用中也是非常难以处理的。[21]

研究非线性时间序列的另一种方法是模型分析法。模型的形式是多种多样的，主要有可加噪声模型、随机条件方差模型、线性/非线性自回归滑动平均模型，以及对上述模型新的组合而得到的一些新的模型，如双重线性模型、非线性多元时间序列模型等。非线性模型不像线性模型那样有统一的表达方式，不存在一种包括所有非线性模型而又便于理论和统计处理的统一形式，即使对有限参数的非线性模型也不能做到这一点。因此，对非线性时间序列只能针对不同类型的模型分别进行研究。某些类型的模型在某些领域中有较好的应用前景，而对另外的领域却不适合，这也是非线性时序分析中的难题之一。[21]

1.4.2 非线性方法

如果知道描述一个系统的数学表达式，那么产生和分析它的时间序列是比较容易的，也不难判定一个时间序列是不是混沌的。由于已经知道系统的变量数，也可以直接构造它的相空间和吸引子，并计算吸引子的分维数或一个轨道的Lyapunov[23]指数。然而，如果我们要研究一个既不知道准确的数学描述方法，又不知道描述变量的准确数目，甚至无法很好控制的系统，例如，大气、流行病、地震、股票市场等，上述的任务变得非常困难。近30年来，

Takens,Grassberger,Procacci,Broomhead,Casadagli,Farmer 等一批科学家发展了一系列分析方法,包括相空间重构、混沌轨道各统计参数的计算以及直接从时间序列数据构造动力学模型并进行非线性预测的技术和算法。[19, 20, 24]

借助这些方法,对这类非线性系统,可以通过对时间序列的分析研究,提取尽可能多的有用信息,识别系统的非线性特征,判断潜在的动力学模型,并对系统的未来行为进行定量的模拟和预测。

本节简要地介绍了非线性时间序列分析的一些基本概念,并对相空间重构以及非线性时间序列的预测等问题做了简单的介绍。如果希望进一步了解这一方面的内容,可以参阅本章末的文献及其相关参考文献。

1.4.3 从观测结果重构系统动力学

对于一个观测时间序列,如果确信它是由一个确定性系统产生的,是否可能根据观察到的数据在一定程度上对系统进行重构呢?

Takens[24]指出,系统中任一分量的演化都是由与之相互作用着的其他分量所决定,因此,这些相关分量的信息就隐含在任一分量的发展过程之中。Packard[25]等人提出的时间延迟的思想,可重构出动力学系统的相空间,这对于不能直接测量系统深层的自变量而仅仅知道一组单变量时间序列的研究人员来说,也有了研究系统动力学行为的可能。例如,已知地面气压观测数据,可通过重构反演出部分高空信息,弥补高空观测资料的不足。它的基本思想是:系统中相关分量的信息隐含在任一分量的发展过程中,为了重构一个"等价"的状态空间,只需考察一个分量,并将它在某些固定的时间延迟点(比如一秒前、两秒前等)上的测量值作为新维处理,即延迟值被看成是新的坐标,它们确定了某个多维状态空间中的一点状态。重复这一过程并测量不同时间的各延迟量,就可以产生出许多这样的点,然后再运用其他方法来检验这些点是否存在于同一个混沌吸引子上。

假设在某动力学系统中,唯一可观察到的是单变量时间序列$\{x_i\}$,为了研究这个时间序列的动力模型,必须重构相空间。这个时间序列的过去状态含有现在状态的信息,这个信息可表示为延迟向量:

$$X(t)=(x(t),x(t-\tau),\cdots,x(t-(m-1)\tau))^T \tag{1.20}$$

其中m为嵌入维数,τ为延迟时间。利用过去状态构造现在状态是非线性时间序列分析的一个标准做法。经过相空间重构,一些不变量如分数维、Lyapunov 指数等得到保留。相空间重构可把具有混沌特性的时间序列重建为一种低阶非线性动力学系统,它是非线性时间序列分析的重要步骤,重构的质量直接影响到模型的建立和预测。延迟时间τ和嵌入维数m的选择是相空间重构的两个重要参数。[26~27]

嵌入维数是动力学系统吸引子的一个重要特征,它定义为描述吸引子上一个点的位置而需要的独立坐标分量个数,从而确定该动力学系统的有效自由度数目。吸引子的维数通常小于嵌入空间维数。可以证明,任何一个m维的光滑流形可以光滑地嵌入到维数为$s=2m+1$的相空间中,而且选择一个大于$2m+1$维的相空间去嵌入,一定可以保留吸引子的拓扑性质。在实际操作过程中,通常难以预先确定嵌入维的维数,只能采用试探的方法。[27]

相空间重构方法对时间延迟τ的选择具有高度敏感性。如果τ太小,则所构造的吸引