应用混合分布研究银川平原地下水埋深对植被的影响

孙宪春， 金晓媚， 万 力

中国地质大学(北京) 水资源与环境学院，北京 100083

Sun Xianchun， Jin Xiaomei， Wan Li

School of Water Resource and Environment, China University of Geoscience (Beijing), Beijing 100083, *China*

Sun Xianchun, Jin Xiaomei, Wan Li. The effect of groundwater level on the vegetation from the mixture of distributions in Yinchuan plain, northwest China. *Earth Science Frontiers*, 2008, 15(5): 344-348

Abstract: The condition of natural eco-environment in Yinchuan plain located in the arid northwestern area is typical in China. MODIS NDVI is one of the important indexes that represent the condition of natural eco-environment, which is influenced by various factors. Groundwater is the most important one among the factors, but it is difficult to make full use of data concerning groundwater to establish a mathematical model and to draw a quantitative conclusion. Currently, the study on the mixture of distributions has received widespread attention, but the reliable estimation of the parameters in the mixture of distributions is quite difficult. A method, based on the relation between the mixture of distributions and the histogram of data, and using least squares to estimate the parameters, is introduced in this paper. The relation between groundwater and vegetation in Yinchuan Plain is discussed on the basis of the new method, using MODIS NDVI and groundwater monitoring data. The normal distribution of the effect of groundwater on vegetation in Yinchuan plain is founded.

Key words: mixture of distributions; least squares; histogram of data; NDVI; groundwater

摘 要:银川平原地处西北干旱地区，自然生态环境在西北地区具有典型性。研究银川平原的生态环境，需要研究植被，它反映多种因素的作用。在影响植被的各种因素中，地下水对植被生长的影响非常重要，但将地下水对植被生长的影响进行定量化分离并不容易。近年来，用混合分布函数处理大量数据的方法受到越来越多的关注，但是混合分布函数的参数求解一直比较困难。文章提出了一种基于混合分布函数和数据直方图之间的关系，用最小二乘法估计参数的新方法；并用这种方法，结合银川平原遥感数据中的归一化植被指数 NDVI，对地下水埋深与植被生长的关系进行定量分析，得到地下水埋深对植被影响所服从的正态分布密度函数。

关键词:混合分布；最小二乘法；直方图；NDVI；地下水

中图分类号:P641.1;P641.8 **文献标识码**:A **文章编号**:1005-2321(2008)05-0344-05

植被是联结土壤、大气和水分的自然“纽带”。植被的生长状况可以用来衡量一个地区环境状况的好坏。目前，遥感数据中的植被指数是反映区域性植被状况的理想数据，它以定量的方式合理地表示地面植被覆盖程度。该指标表征地面范围的大小则取决于遥感数据的分辨率，如 NOVA 卫星数据的一

收稿日期:2007-10-10;**修回日期**:2008-05-15

基金项目:国家自然科学基金项目(40772158)

作者简介:孙宪春(1972—)，男，博士研究生，从事水文地质方面的研究。E-mail:sxc6688@126.com

个数值反映 1 km² 范围的植被覆盖程度。当进行区域性大范围的植被状况定量描述时,由于数据量较大,影响因素众多,统计学方法进行分析是必要的。然而,当有两个以上因素对植被生长状况产生影响时,植被指数的统计分布往往呈现出两个甚至多个正态分布的混合。只有将这些混合正态分布进行分离,才能深入了解各种自然因素与植被生长之间的关系。

混合分布函数的研究始于 20 世纪 50 年代。Davis[1]在研究元件寿命试验中提出污染变量所服从的混合分布形式,认为元件的寿命分布函数为两个具有同样形式的分布函数混合,并构造了相应的分布函数。Huber[2]考虑一类"被污染的正态分布族",在标准正态分布的基础上,叠加了一个关于原点对称的分布。Quandt[3]讨论了二正态混合分布的参数估计问题。From[4]讨论了二指数分布的混合分布问题。郑祖康等人[5]对于两正态混合分布在实际工程中的应用进行研究,并讨论了混合正态分布参数的矩估计方。目前,混合分布函数应用的关键问题是如何估计分布中的参数以及如何在实际工程中结合数据进行分析,因为很多自然现象是多种因素作用的综合表现,分离出各个因素的具体影响就需要对其服从的方程求解,给出定量化的结论,然而混合分布函数的参数并不容易求得。

目前,混合分布中参数求解主要有矩估计法,最大似然估计法,贝叶斯估计法等[6],这些参数估计方法,对于参数较少的情况,相对容易求解;但是对于多参数的情况就很困难,有时甚至不可解;尤其当方程含有 5 个以上的未知数时,建立求解方程组本身是困难的,即使建立方程组,也只能通过数值的方法求解。在数值解法中,收敛解常常与选取的初始值有关,有时因为初始值选取的不合适会导致方程不收敛;尤其是高次方程组,收敛的初始点是很难给定的。

在实际应用中,人们常常用直方图来直观分析样本数据的分布特性,这种方法直观、简单,但是因为多因素作用的大样本数据在直方图中表现为不规则的单峰分布、双峰分布或者是多峰分布(见图 1,3,4),特性复杂,所以不能定量给出混合分布的参数估计值。

为了克服建立、求解多元高次方程组的困难以及弥补用直方图不能进行定量分析的不足,本文讨论了一种求解混合分布参数的新方法,来满足工程

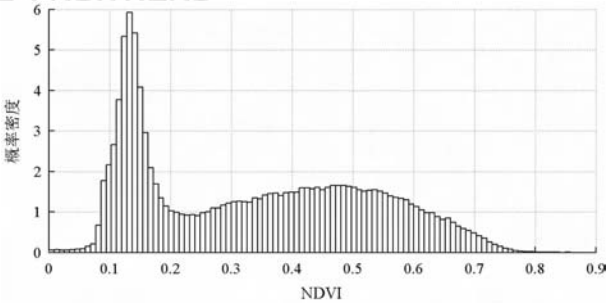

图 1　NDVI 直方图

Fig. 1　The histogram of NDVI

中进行定量分析的需要,即混合分布函数与累积频率直方图拟合法,当混合分布函数连续时即为混合分布密度曲线与频率密度直方图拟合法。

1　混合分布函数的参数求解方法

一般说来,混合分布函数具有以下形式:

$$F(x)=\sum_i \alpha_i F_i(x) \tag{1}$$

其中,x 为随机变量;F_i 为各因素所服从的分布函数;α_i 为混合系数,反映各分布在总体分布中占的比例,α_i 越大,则该因素的影响越大;α_i 的取值范围 0～1,$\sum_i \alpha_i=1$。

当 F_i 为正态分布时,

$$F_i(x,\mu_i,\sigma_i)=\frac{1}{\sqrt{2\pi}\,\sigma_i}\int_{-\infty}^{x}\exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]\mathrm{d}x \tag{2}$$

其中,μ_i 为正态分布的数学期望;σ_i 为正态分布的均方差。

当建立了混合分布函数后,具体求解方法如下:(1)将 x 均分成 i 个$[x_{i-1},x_i]$的子区间;(2)利用直方图求出累积频率 $Y_i=\frac{\sum \Delta n_i}{n}$,$\Delta n_i$ 为 Δx 区间内 x 的个数,n 为 x 的总个数;(3)将 x_i 的值代人方程 $\min \frac{1}{2}\sum_i (F(x_i)-Y_i)^2$ 求到稳定的收敛解。

当混合分布函数连续时,为了简化计算:(1)建立混合分布密度函数 $f(x)$,将 x 的值均分成 i 个$[x_{i-1},x_i]$的子区间;(2)利用直方图求出频率密度 $y_i=\frac{\Delta n_i}{n\Delta x}$值,$\Delta n_i$ 为 Δx 区间内的 x 的个数,n 为 x 的总个数,$\Delta x=x_i-x_{i-1}$;(3)将 x_i 的值代人方程 min

$\frac{1}{2}\sum_i (f(x_i)-y_i)^2$求到稳定的收敛解。

2 用混合分布研究银川平原地下水埋深对植被的影响

银川平原地处西北干旱地区,是中国黄河流域重要的工农业生产基地,自然生态环境在西北地区具有典型性,对其生态环境的研究具有重要意义。研究银川平原的生态环境,需要研究植被,它是反映生态环境状况的重要指标之一,它反映多种因素的综合作用,在各种影响因素中,地下水对植被生长的影响非常重要。针对地下水在植被生长中的作用,有些学者进行了研究[7-9]。在银川平原,地下水埋深在 3 m 左右时,植被长势最好;适宜于植被生长的地下水埋深 1～5 m;地下水埋深超过 5 m,则对植被的影响逐渐减弱;当地下水埋深小于 8 m 时,则其影响可以不计。但是,这些研究得到的只有定性的分析,没有给出定量化的结论。本文的目的是应用正态混合分布,将银川平原地下水对植被生长的影响进行定量分析。由于本文研究的是大尺度问题,采用的是遥感数据中的归一化植被指数(NDVI),并且对 NDVI 根据地下水埋深进行了几何平均处理,所以认为在研究区域除了地下水的影响作用不同以外,其他各种因素对 NDVI 的作用都是相同的,即认为地下水埋深小于 8 m 时,影响 NDVI 的因素是地下水的埋深以及其他因素,当地下水埋深大于 8 m 时,只有其他因素的影响。

本文采用的是 16 d 合成的,空间分辨率为 250 m,可以用来检测地球植被的季节变化和年际变化的 MODIS NDVI 数据,根据实测的 2004 年 4 月的地下水埋深数据,通过插值得到与 MODIS NDVI 分辨率一致的地下水埋深网格数据,将地下水埋深和 NDVI 数据相对应,在研究区内同一位置上分别得到地下水埋深和 NDVI 值。为了分析地下水埋深与植被生长状况的关系,我们取地下水埋深相同的数据点所对应的 NDVI 的均值,代表该地下水埋深条件下的植被长势。

为了考察 NDVI 的分布特性,作直方图(图 1),是不规则的双峰分布,因此首先构造双正态的混合分布函数来初步分析 NDVI 的影响因素。

构造的双正态混合分布函数具有以下形式:

$$F=\alpha F_1+(1-\alpha)F_2 \tag{3}$$

F_1,F_2 的意义见式(2)。

用公式(3),将 x_i 和 Y_i 的值带入方程 min $\frac{1}{2}\sum_i (F(x_i)-Y_i)^2$,用信赖域与预处理共轭梯度法求解[10],在数值求解中,解的误差限为 10^{-4},函数的误差限为 10^{-4},求得收敛的混合系数 α(图 2),其中地下水埋深 7 m 以上按 0.5 m 间距计算,7 m 以下按 2 m 间距计算。

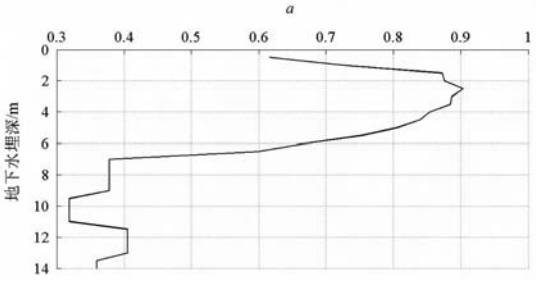

图 2 混合系数 α 与地下水埋深的关系曲线
Fig. 2 The relation between mixture coefficient α and groundwater level

从图 2 看出,当地下水埋深大于 8 m 时,α 在 0.35 左右波动,变化幅度比较小,两个分布的比例稳定;当地下水埋深小于 8 m 时,α 的变化大,需要考虑更多因素的影响。因此,以地下水埋深 8 m 为界限,将 NDVI 的影响因素分为有地下水影响与没有地下水影响两种状态,通过这两种状态的分析,确定地下水埋深对 NDVI 的定量影响。

首先,将地下水埋深大于 8 m 的 NDVI 作直方图(图 3)。从图 3 中可以看到,NDVI 为不规则的单峰分布,本文采用双正态混合分布函数来描述样本数据的主要特性,考虑到正态分布函数的连续性,用其概率密度函数来估计最优参数。构造的密度函数:

$$f=\sum_{i=1}^{2} a_i f_i(x,\mu_i,\sigma_i)=\sum_{i=1}^{2} a_i \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right] \tag{4}$$

其中,$\sum_{i=1}^{2} a_2=1$。将 x_i 和 y_i 的值代入方程 min $\frac{1}{2}\sum_{i=1}^{100}[f-y_i]^2$,求解方法同式(3),收敛的参数值如下:

f_1:$\mu_1=0.235\ 1$,$\sigma_1=0.124\ 3$,$\alpha_1=0.305\ 3$;

f_2:$\mu_2=0.134\ 9$,$\sigma_2=0.021\ 5$,$\alpha_2=0.694\ 7$。

将求解的参数值带入式(4),并与 NDVI 数据直方图叠加,如图 3 所示。

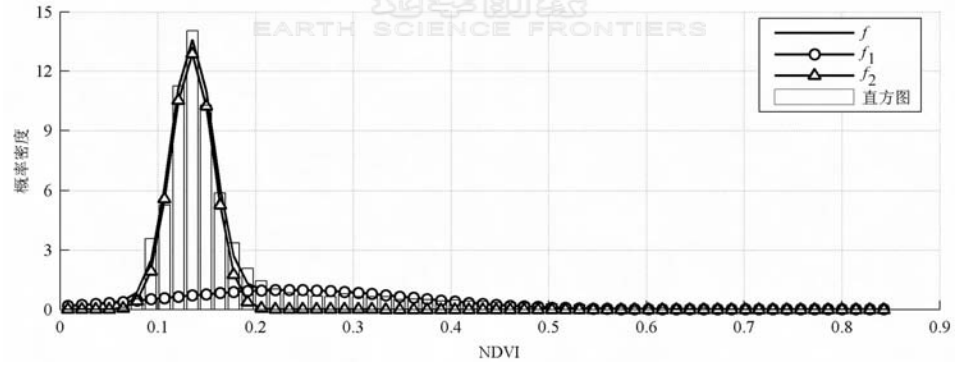

图 3　拟合的密度曲线和 NDVI 直方图

Fig. 3　The fitted density curves and the histogram of NDV

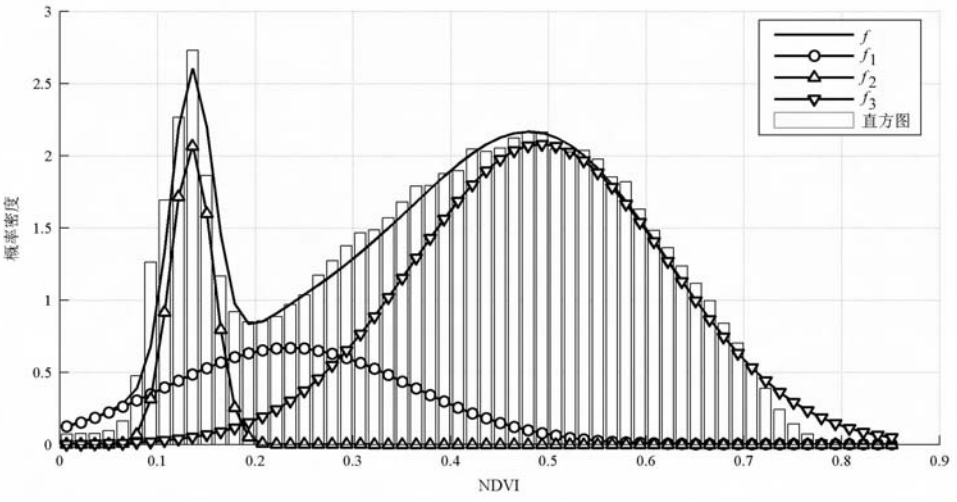

图 4　拟合的密度曲线和 NDVI 直方图

Fig. 4　The fitted density curves and the histogram of NDVI

由图 3 可见，拟合的密度函数反映了样本数据的主要分布特性，在地下水埋深变化对 NDVI 变化影响可以不计的状态下，形成了两因素的混合分布密度函数，第一个以 0.235 1 为中心；第二个以 0.134 9 为中心。

为了分析地下水埋深对 NDVI 变化的影响，作地下水埋深小于 8 m 的 NDVI 直方图(图 4)。从图 4 可以看到，NDVI 的分布为不规则的双峰分布，因此本文构造三正态的混合分布密度函数：

$$F=\sum_{i=1}^{3}a_i f_i(x,\mu_i,\sigma_i)=\sum_{i=1}^{3}a_i\frac{1}{\sqrt{2\pi}\,\sigma_i}\exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right] \quad (5)$$

其中，$\sum_{i=1}^{3}a_i=1$。将 x_i 和 y_i 的值带入方程 $\min\frac{1}{2}\sum_{i=1}^{100}[F-y_i]^2$，求解方法同式(3)，得到收敛的参数值如下：

f_1：$\mu_1=0.235\ 1$，$\sigma_1=0.124\ 3$，$\alpha_1=0.208\ 3$；

f_2：$\mu_2=0.134\ 9$，$\sigma_2=0.021\ 5$，$\alpha_2=0.111\ 5$；

f_3：$\mu_3=0.492\ 3$，$\sigma_3=0.130\ 5$，$\alpha_3=0.680\ 2$。

将求得参数值带入式(5)，并与 NDVI 数据直方图进行叠加，如图 4 所示。

由图 4 可见，拟合的三正态概率密度函数反映了样本数据的主要分布特性，在地下水埋深变化对 NDVI 变化影响明显的状态下，形成了三因素的混合分布密度函数。第一个以 0.235 1 为中心；第二个以 0.134 9 为中心；第三个以 0.492 3 为中心。通过图 4 与图 3 的比较可以看出，NDVI 服从的三因素正态分布中有两个因素与地下水埋深对 NDVI 变化没有影响状态下的分布相同，因此第三个因素的分布是由地下水的作用引起的，即在银川平原地下水埋深对植被 NDVI 的影响服从的分布密度

函数为:

$$f_3=\frac{1}{0.130\ 5\times\sqrt{2\pi}}\exp\left[-\frac{(x-0.492\ 3)^2}{2\times0.130\ 5^2}\right]$$

3 结论

(1)用混合分布函数,结合遥感数据,得到银川平原地下水对植被影响所服从的分布密度函数,定量描述了地下水埋深对植被覆盖度的影响。

(2)用直方图的频数值与概率分布函数的关系,通过最小二乘法求解分布函数参数的方法,适用于多参数分布函数的情况。

(3)混合系数 α_i 反映了在混合分布中,各因素的重要程度。可用 α_i 稳定的数值点作为样本数据不同状态的分类依据。

(4)本文提出的方法为用混合分布函数定量处理工程中的大样本问题提供了参考;在应用中,要结合实际经验,在建模及其实践意义的分析中更有效率。

References:

[1] Davis D J. An analysis of some failure data[J]. JASA, 1952, 47: 113-150.

[2] Huber P J. Robust estimation of a location parameter[J]. Annals of Mathematical Statistics, 1964, 35: 73-101.

[3] Quandt R E, Remsey J B. Estimating mixture of normal distributions and switing regressions[J]. JASA, 1978, 83: 730-738.

[4] From S G. Optimal spacing of quantities for the estimation of the parameters in mixture of two exponential distributions [J]. Commun Statist(Theory Methods), 1989, 18: 2201-2223.

[5] Zheng Z K, Ma R, Chen H Y. The study of the distribution of the Chinese BMI[J]. Mathematical Theory and Applications, 2000, 12(3): 121-128(in Chinese).

[6] Zheng Z K, Wu X M, Rao G. The treatment of contaminated data[J]. Chinese Journal of Applied Probability and Statistics, 1998, 14(3): 307-312(in Chinese).

[7] Jin X M, Xue Z Q, Yu Q S, et al. Groundwater resources development and variation in vegetation and in the Yinchuan plain[J]. Hydrogeology and Engineering Geology, 2007, (3): 33-36(in Chinese).

[8] Jin X M, Wan L, Zhang Y K, et al. A study of the relationship between vegetation growth and groundwater in the Yinchuan plain[J]. Earth Science Frontiers, 2007, 14(3): 197-203(in Chinese).

[9] Sun X C, Jin X M, Wan L. The study on the effect of groundwater on vegetation growth in Yinchuan plain[J]. Geoscience, 2008, 22(2): 143-146(in Chinese).

[10] Chen B L. Optimal theory and arithmetic[M]. 2nd ed. Beijing: Tsinghua University Press, 2006: 315-328(in Chinese).

参考文献:

[5] 郑祖康,马蓉,陈汉元. 关于我国体质指数 BMI 的分布研究[J]. 数学理论与应用, 2000, 12(3): 121-128.

[6] 郑祖康,吴雪明,饶刚. 污染数据的处理[J]. 应用概率统计, 1998, 14(3): 307-312.

[7] 金晓媚,薛忠歧,余秋生,等. 银川平原地下水资源开发与植被变化[J]. 水文地质工程地质, 2007, (3): 33-36.

[8] 金晓媚,万力,张幼宽,等. 银川平原植被生长与地下水关系研究[J]. 地学前缘, 2007, 14(3): 197-203.

[9] 孙宪春,金晓媚,万力. 地下水对银川平原植被生长的影响研究[J]. 现代地质, 2008, 22(2): 143-146.

[10] 陈宝林. 优化理论与算法(第 2 版)[M]. 北京: 清华大学出版社, 2006: 315-328.