

环境量缺失数据处理方法探讨

徐 波^{1,2,3}

(1. 河海大学水利水电学院,江苏 南京 210098 ;2. 河海大学水文水资源与水利工程科学国家重点实验室,江苏 南京 210098 ;
3. 河海大学水资源高效利用与工程安全国家工程研究中心,江苏 南京 210098)

摘要 介绍了几种环境量缺失观测数据的处理方法,具体包括样条插值函数、相似年法、蒙特卡洛法。用这些方法补齐了某重力坝的环境量缺失数据,并用处理后的数据建立统计模型,发现上述方法可以较好地补齐环境量的缺失数据,而且环境量观测数据处理后所建模型精度有一定提高。
关键词 大坝安全监控 环境量 缺失数据 数据补齐方法 样条插值函数 相似年法 蒙特卡洛法
中图分类号 TV698.1 **文献标识码** A **文章编号** 1006-7647(2010)S1-0035-03

大坝安全监控中,对监测资料进行定量的分析,主要是建立效应量和环境量之间的数学模型,例如统计模型、确定性模型、混合模型等并由此定量分析评价大坝的安全状态^[1]。在建立模型前,为更准确地确定环境量和效应量之间的关系和提高模型的精度,必须对观测数据进行处理。观测数据中存在突变、噪声、甚至错误的数据,同时由于一些原因,环境量的观测数据可能缺失。鉴于此,如果不对观测数据进行处理,直接采用观测数据建立模型,所建模型的精度较低,拟合效果较差,不能准确反映各环境量对效应量的影响程度,从而也不能合理的分析评价大坝的安全状态,不能对效应量的发展趋势做出准确的预报。观测数据处理主要包括去掉大的突变、降低噪声的影响、缺失数据处理以及均值计算等。本文主要针对环境量的缺失数据处理的方法进行探讨。一般来说,环境量的时间序列是等间距的,但由于某时段数据未观测,观测数据记录丢失,观测数据可靠性较差、不宜采用等多种原因,使得有些环境量观测数据缺失,其时间序列不满足等间距要求。此时,可以将环境量缺失数据采用样条插值函数、相似年法、蒙特卡洛法等方法补齐。

1 环境量缺失数据处理方法

根据缺失数据的多少,环境量缺失数据的处理方法可以分为 2 种:较少数据补齐方法和较多数据补齐方法。假定观测数据每天都有记录,对缺失数

据较少的情况(n 小于或等于 7, n 为每月缺失数据个数)通过较少数据补齐方法处理缺失的数据;对缺失数据较多的情况(n 大于 7)则采用较多数据补齐方法处理缺失的数据。

1.1 较少数据补齐方法

对缺失数据较少的情况,采用三次样条插值函数补齐缺失数据。三次样条插值函数具有二阶连续导数,即其曲线具有二阶光滑度,而且三次样条插值函数的误差较小^[2]。下面介绍三次样条插值函数的原理及其应用方法。

假设在区间 $[a, b]$ 上给定 $n + 1$ 个插值点 $a = x_0 < x_1 < \cdots < x_n = b$ 及其函数 $f(x)$ 对应值 y_i ,若函数 $S(x)$ 满足:① $S(x)$ 在每个子区间 $[x_j, x_{j+1}]$ ($j = 0, 1, 2, \cdots, n - 1$) 是一个三次多项式;② $S(x)$ 在每个内节点 x_j ($j = 0, 1, 2, \cdots, n - 1$) 具有直到二阶的连续导数,即 $S(x) \in C^2[a, b]$,则称 $S(x)$ 为节点 $x_0, x_1, \cdots, x_n$ 的三次样条函数。

如果 $S(x)$ 在满足条件(1)(2)的同时还满足 $S(x) = y_i$ ($j = 0, 1, 2, \cdots, n - 1$),则称 $S(x)$ 为三次样条插值函数。

在用三次样条插值函数补齐环境量缺失数据时,根据缺失数据的时间划分样条插值函数的子区间 $[x_j, x_{j+1}]$,并借助 MATLAB 样条插值工具箱提供的 csapi() 函数来求得三次样条插值函数^[3],从而求得缺失的数据。

1.2 较多数据补齐方法

对缺失数据较多的情况,采用相似年法^[4]以及蒙特卡洛法^[5]的相关原理进行处理。相似年法可以根据环境量周期变化的特点补齐较多的缺失数据,使得环境量测值等间距、连续分布。由于降雨量测值连续性较差,存在一定的离散性,而且降雨量的随机性比较强,在此对于降雨量测值缺失较多的情况借助蒙特卡洛法的伪随机数和抽样模拟^[5-6]来补齐缺失的数据。

1.2.1 相似年法

相似年法是在缺失数据较多的情况下,找到与缺失数据年份测值相似的年份,并将缺失数据用相似年份的相应测值代替的数据处理方法。相似年法可以根据环境量周期变化的特点补齐较多的缺失数据,使环境量测值等间距、连续分布。

确定相似年的具体方法为:用矩法^[7]求得各年份测值的均值($\bar{X}$),变差系数(C_v),若某年测值的 $\bar{X}$ 及 C_v 与缺失数据年份的相近,则把该年份看作相似年,用其相应的数据代替缺失数据。计算均值、变差系数采用的具体公式分别为

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i \tag{1}$$

$$C_v = \frac{S}{\bar{X}} \tag{2}$$

式中: $\bar{X}$ 为均值; X_i 为观测值; N 为测值个数; C_v 为变差系数; S 为标准差, S 由公式 $S^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2$ 确定。

1.2.2 蒙特卡洛法

蒙特卡罗法的基本原理为:首先依据所求的问题建立所求问题解的基本表达式,确定出基本变量和基本变量的概率模型,然后,用计算机方法生成伪随机数,根据伪随机数对模型抽样试验来计算问题的解。在此借助蒙特卡洛法中伪随机数的产生和抽样模拟来补齐降雨量缺失的观测数据。

1.2.2.1 伪随机数的产生

伪随机数是在计算机上通过数学迭代过程产生的一系列数,并通过局部随机性如均匀性、独立性等检验。伪随机数要具有以下优点:随机性好、周期长、在计算机上容易实现且节约机时等。产生伪随机数有以下几种方法:平方取中法、位移指令加法、乘同余法、加同余法以及混合法等。在此采用各领域中广泛应用的乘同余法来产生伪随机数,该法的迭代公式为

$$X_{i+1} = \lambda X_i \pmod{M} \tag{3}$$

$$r_i = X_i M^{-1} \tag{4}$$

式中: $M = 2^s$,取 $s = 32$, $X_0 = 1$, $\lambda = 5^{13}$, X_0 要随机选取一个 $4q + 1$ 型的数, q 为任意整数, λ 取成 5^{2k+1} 型的正整数,其中 k 为使 5^{2k+1} 在计算机上所能容纳的最大奇数。

1.2.2.2 随机变量的抽样模拟

确定了基本变量的概率模型后,可以根据产生的一系列随机数进行抽样。根据文献^[8],随机变量分布的模拟方法在水工建筑物安全监测领域常见的有正态分布、对数正态分布以及极值Ⅰ型分布等。在此假定每年同一天的降雨量满足对数正态分布,则缺失数据时段每天的降雨量可由下式确定:

$$X_i = \exp r_i \quad (i = 1, 2, \cdots, n) \tag{5}$$

式中: X_i 为抽取的每天降雨量值; r_i 为由式(4)产生的伪随机数, r_i 服从正态分布。

1.2.2.3 抽样次数的确定

模拟计算所需要的抽样次数是由研究对象的性质和特点决定的。但是在实际运用中,可以根据具体研究对象,人为设置模拟抽样次数。目前常用的主要有以下2种:①人为设定最大试验次数 N ,当模拟抽样次数达到 N 时便停止计算;②每计算 n 次,便将计算结果与前一个 n 次的结果做对比,如果差值满足一定的条件则停止计算,依此来获得计算分析的总试验次数。在此参照第1种方法,设定抽样的次数为10次,将10次抽样结果的均值作为缺失数据时段每天降雨量的替代值。

1.3 缺失数据处理方法的程序流程

根据上述原理,编写了环境量缺失数据处理程序,用该程序处理环境量观测数据,并用处理后的数据建立统计模型,考察模型精度,以模型精度的差异来说明缺失数据处理的效果和必要性。环境量缺失数据处理程序的主要结构框架见图1。

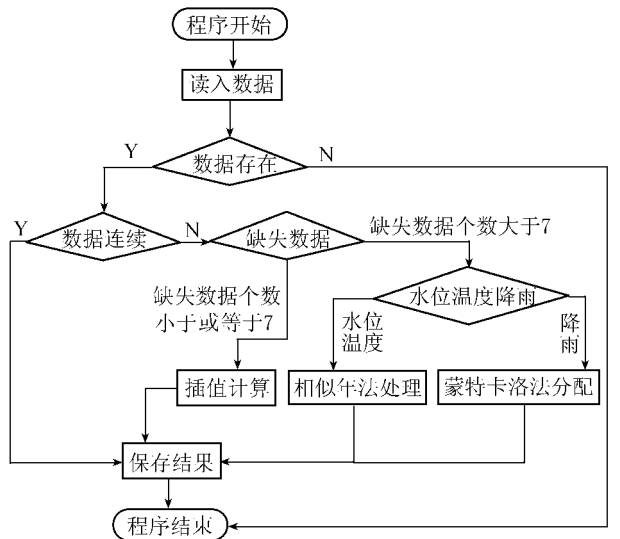


图1 环境量缺失数据处理程序

2 应用实例

2.1 工程概况

某水利枢纽的主要挡水建筑物为一重力坝,坝顶高程为 78.00 m,最大坝高为 51.5 m,总库容为 7468 万 m³,坝址上游集水面积为 9 000 km²。现有该坝 1992—2007 年共 15 年的监测资料,总体上监测资料较完善,缺失数据较少,笔者选择 2000 年和 2001 年的环境量监测资料作为研究对象。2000 年 5 月份上游水位和降雨的部分观测数据缺失,缺失数据个数均小于 7。2001 年 4 月份缺失了上游水位的全部观测数据和 15 d 降雨的观测数据,缺失数据个数分别为 30 和 15,大于 7。用环境量缺失数据处理程序对上述缺失数据处理,并计算环境量均值。选择 2000—2001 年 2 年的观测资料建立统计模型,并比较实际观测数据所建立模型与程序处理后的数据所建立模型的精度。

2.2 计算成果

2000 年 5 月份的观测数据,上游水位缺失了 5 个观测数据,缺失观测数据的日期分别为 2000 年 5 月 10—11 日、18 日、27 日、30 日。降雨缺失了 3 个观测数据,缺失数据的日期分别为 2000 年 5 月 22—23 日、29 日。上游水位和降雨的观测数据由程序插值计算后的情况由图 2、图 3 给出。图 2 和图 3 为补齐缺失数据后 2000 年 5 月份的上游水位和降雨过程线。

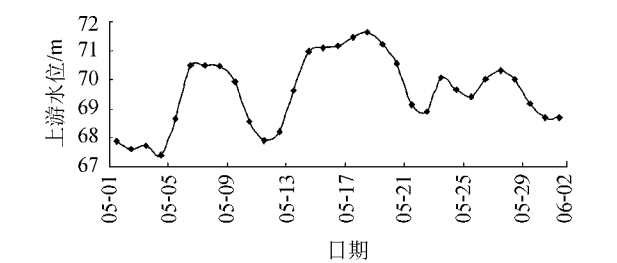


图 2 程序处理后的 2000 年水位过程线

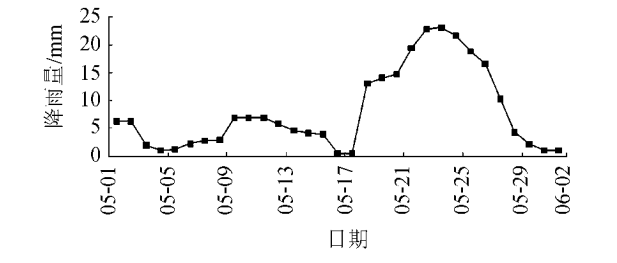


图 3 程序处理后的 2000 年降雨过程线

由于一般情况下环境量的测值呈等间距、连续性变化,由图 1、图 2 可以看出,补齐数据处曲线的连续性和光滑度较好,说明样条插值函数可以用来解决缺失较少数据的补齐问题。需要指出的是降雨量的连续性较差,用样条插值函数补齐的缺失数据存在一定的误差。

2001 年 4 月份的观测数据缺失了上游水位的全部测值,用相似年法找到与该年上游水位相近的年份,并将 2001 年 4 月份缺失的上游水位的观测数据用相似年份的数据代替。相似年查找时的计算数据如表 1 所示。

表 1 上游水位的均值、变差系数统计

年份	水位 均值/m	水位变差 系数	年份	水位 均值/m	水位变差 系数
1992	59.99	0.10	2000	69.87	0.02
1993	69.38	0.03	2001	69.40	0.03
1994	69.00	0.03	2002	69.63	0.02
1995	69.25	0.03	2003	69.83	0.01
1996	68.69	0.04	2004	68.92	0.02
1997	69.87	0.03	2005	69.10	0.03
1998	70.09	0.02	2006	70.52	0.01
1999	69.91	0.02	2007	70.36	0.02

由表 1 可以看出:2001 年上游水位的均值为 69.40 m,变差系数为 0.03。与 2001 年上游水位均值和变差系数都相近的年份是 1993 年,1993 年上游水位的均值为 69.38 m,变差系数为 0.03,因此选择 1993 年作为 2001 年上游水位测值的相似年,将 2001 年 4 月份缺失的上游水位观测资料用 1993 年 4 月份的代替。

2001 年 4 月份缺失了 15 d 的降雨观测资料,假定每天的降雨均服从对数正态分布,按蒙特卡洛法模拟抽样 10 次并取均值后求得每天的降雨量分别为 7.12 mm,6.59 mm,7.21 mm,3.77 mm,3.96 mm,3.00 mm,2.65 mm,8.01 mm,1.24 mm,3.57 mm,4.27 mm,1.89 mm,2.64 mm,2.05 mm,2.97 mm。

选取坝基渗漏量作为效应量,由实际观测的环境量数据和由程序处理后的环境量数据分别建立统计模型,表 2 给出了坝基渗漏量测点 WE[A]01 统计模型在数据处理前后的复相关系数 R 和剩余标准差 S 。

表 2 数据处理前后统计模型 R 、 S 对比

实际数据		处理数据	
统计模型 R	统计模型 S	统计模型 R	统计模型 S
0.74	0.58	0.81	0.35

由表 2 可看出:人为处理后的数据所建立模型的复相关系数 R 与实际比较有一定的增大,由 0.74 增大为 0.81。剩余标准差 S 与实际相比有所减小,由 0.58 减小为 0.35。总体上用该程序处理的数据所建统计模型能够保证模型的精度,并且使统计模型的精度有所提高。

3 结 语

环境量观测数据处理是大坝安全监控中建立模型前的重要准备工作,对建模效率和(下转第 47 页)

利用

For Y = 0 To Picture1.Height Step 1'纵向扫描

For X = 0 To Picture1.Width Step 1'横向扫描

ForeColor = Picture1.Point(X,Y)'取得扫描点的颜色像素值

逐像素点扫描比较,最后结果通过 MsgBox 函数创建的消息框来实现,见图 4。



图 4 MsgBox 函数消息框

通过对抽取出的大量位图图像的识别计算,验证了以图像灰度标准差这一参数进行相同图像识别剔除的方法是可行的,并且和其他识别方法^[2-6]相比,前者更高效,从而为微小气泡研究的前期预处理工作节省了大量时间。

4 结 语

本文提出一种以图像灰度标准差为判别标准、基于 VB 应用程序的自动识别方法,来处理将视频图像抽取成平面位图图像过程中出现的相同图像的问题。通过对大量图像的运算论证,证实了该方法处理结果准确,耗时较少,在实际工程应用中有一定的可行性。

参考文献:

- [1] 张蓉生,刘志鹏,李春晓.平面图像里重叠微小气泡的分离与测量[J].流体力学,2006,27(1):38-39.
- [2] 甘进,王晓丹,权文.基于特征点的快速匹配算法[J].电光与控制,2009,16(2):64-66.
- [3] THORNTON J, SAVVIDOS M. A Bayesian approach to deformed pattern matching of iris image [J]. IEEE Pattern Analysis Machine Intelligence, 2007, 29(4): 596-606.
- [4] 陆欢,吴庆宪,姜长生.多传感器图像的二次匹配算法[J].电光与控制,2008,15(1):22-25.
- [5] 彭肇,刘晓平.基于内容的可移植图像识别系统研究[J].计算机与数字工程,2010,38(1):140-142.
- [6] 孙红辉,姚良,张清华,张红蕾.基于神经网络的图像识别方法研究[J].光学技术,2008,34(增刊):173-175.
- [7] ZHANG Rong-sheng, FUJIKAWAS, CHEN Ci-chang. Separation and segmentation for recognition and measurement of overlapped bubbles in plane images [J]. Journal of Information and Computational Science, 2005, 2(4): 815-828.
- [8] FUJIKAWAS, ZHANG Rong-sheng, HAYAMAS. The control of

micro-air-bubble generation by a rotational porous plate [J]. International Journal of Multiphase Flow, 2003, 29(8): 1221-1236.

- [9] 张蓉生,郑源,程云山.微小气泡粒径的测量研究[J].实验流体力学,2005,19(2):91-93.
- [10] 管业鹏,顾伟康.二维图像特征点自主提取算法研究[J].传感技术学报,2004,17(1):71-72.

(收稿日期 2010-06-08 编辑 骆超)

(上接第 37 页)模型精度有较大的影响。本文介绍了环境量缺失数据的处理方法,包括样条插值函数、相似年法、蒙特卡洛方法等,并借助程序的手段用这些方法补齐了某重力坝的环境量缺失数据,用处理后的环境量观测数据建立了统计模型与未处理的数据所建模型进行了对比。通过分析比较可以发现:①样条插值函数、相似年法、蒙特卡洛方法等可以解决环境量缺失观测数据的补齐问题。②用上述 3 种方法处理后的数据所建立的统计模型能够保证模型的精度,并使模型精度有一定程度的提高。将这 3 种方法用程序实现可以提高环境量数据处理的效率,节省建模的准备时间。

在用相似年法和蒙特卡洛法处理数据时有些问题值得进一步研究:①监测资料时间序列较短、年份较少时,相似年如何选取;②降雨量的概率分布如何快速准确地确定。

参考文献:

- [1] 吴中如.水工建筑物安全监控理论及其应用[M].南京:河海大学出版社,2003:6-9.
- [2] 李庆扬,王能超,易大义.数值分析[M].北京:清华大学出版社,2001:51-58.
- [3] 薛定宇,陈阳泉.高等应用数学问题的 MATLAB 求解[M].北京:清华大学出版社,2005:271-275.
- [4] 梁忠民,钟平安,华家鹏.水文水利计算[M].北京:中国水利水电出版社,2006:23-33.
- [5] 徐钟济.蒙特卡罗方法[M].上海:上海科学技术出版社,1985:1-5.
- [6] 谷艳昌,何鲜峰,郑东健.基于蒙特卡罗方法的高拱坝变形监控指标拟定[J].水利水运工程学报,2008(1):14-19.
- [7] 刘国祥,何志芳,杨纪龙.概率论与数理统计[M].兰州:甘肃教育出版社,2002:215-255.
- [8] 沈凤生,吴世伟.求解结构可靠度的蒙特卡洛法及其改进[c]//吴世伟.结构安全度与可靠度分析论文集.南京:河海大学出版社,1987:43-51.

(收稿日期 2009-10-12 编辑 高建群)