DOI:10.3880/j.issn.1006-7647.2018.05.013

·庆贺吴中如院士 80 华诞专刊论文·

基于统计诊断的大坝监测数据合理性检验

李子阳[1],郭 丽[2],马福恒[1],胡 江[1]

(1.南京水利科学研究院水文水资源与水利工程科学国家重点实验室,江苏 南京 210029;
2.南京体育学院附校部,江苏 南京 210024)

摘要:基于统计诊断的异常数据划分,并结合大坝监测数据的误差成因,将监测的异常数据划分为随机误差、粗差、系统误差等,并辨识强影响数据。继而基于均值漂移模型,研究不同异常数据的诊断方法,包括以模型扰动值为依据的粗差的 t 检验法和以模型扰动对拟合参数的影响为依据的强影响数据的 Cook 距离检验法。以大坝典型位移监测数据为例,采用上述统计诊断方法对原始监测数据进行合理性检验,结果表明可有效辨识误差数据和强影响数据,能提高数据进一步建模分析的准确性。

关键词:大坝;监测数据;合理性检验;统计诊断;均值漂移模型

中图分类号:TV698.1 **文献标志码**:A **文章编号**:1006-7647(2018)05-0071-05

Rationality test of dam monitoring data based on statistical diagnosis//LI Ziyang[1], GUO Li[2], MA Fuheng[1], HU Jiang[1](1. *State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering*, *Nanjing Hydraulic Research Institute*, *Nanjing* 210029, *China*; 2. *Accessory School of Nanjing Sport Institute*, *Nanjing* 210014, *China*)

Abstract: According to the abnormal data partitioning of the statistic diagnosis and the causes of the error data in dam monitoring, the abnormal data observed can be categorized into random error, gross error, system error. The influential data is distinguished and then the processing methods for different errors are studied using the mean shift model, including the t-test method for the gross error detection based on the model disturbed values and the Cook-distance-test method for the influential data based on the influence of the model disturbance on the model fitting parameters. Taking typical original monitoring data of dam displacement as an example, rationality test was performed using the proposed statistical diagnosis methods. The results indicate that these methods can effectively distinguish gross errors and influential data, with which the accuracy of further model analysis can be significantly improved.

Key words: dam; monitoring data; rationality test; statistical diagnosis; mean shift model

在大坝安全监测中,受大坝性态演化和作用环境[1]、观测随机因素及仪器本身监测精度等的影响,监测数据不可避免地存在误差[2]。大样本的自动化监测数据,一般存在显著的随机特征,即数据本身存在随机误差;受监测过程中的不确定因素影响,如仪器的不稳定或监测基点发生位移,还可能产生系统误差等。误差的存在影响模型分析的准确性,因此,对大坝监测数据进行合理性检验,以获取更为合理有效的基础分析数据,是监测资料分析和工程性态评估的首要工作。

基于统计分析的假设检验(如 PauTa 准则、t 检验法及 Dixon 判别法等)是监测数据误差检验的常用方法[3-5],对自变量数据(基础环境量数据)的误差分析是有效的,如通过测值范围和方差对传感器数据进行误差检验等[6]。但大坝监测数据受水压、降雨、温度、时效等环境因素的综合影响,监测数据为因变量数据,若只对数据本身或模型参数进行常规的统计检验分析,极有可能会把因环境突变而引起的监测数据改变误判为误差数据,导致有用数据被误删。另外,常规统计方法在分析数据对模型的影响程度和趋势性方面也有所欠缺[7]。

引入统计诊断的方法进行数据的检验分析,可以很好地解决上述问题。统计诊断[8]首先根据因变量和自变量之间的影响关系构建统计模型,进而

基金项目:国家重点研发计划(2016YFC0401602);国家自然科学基金(51779155);中央级公益性科研院所基本科研业务费专项(Y718002)

作者简介:李子阳(1983—),男,高级工程师,博士,主要从事水库大坝健康诊断研究。E-mail:zyli@nhri.cn

借助统计诊断量检查数据、模型及推断方法中可能存在的问题,其在综合考虑大坝监测中自变量数据与因变量数据内在关联性方面具有优势,可为监测数据的合理性检验提供更符合工程实际的方法。统计诊断已在滑坡体位移监测数据异常值检验中有所应用[9],本文在此基础上,从大坝监测数据的异常数据类型分析出发,进一步研究基于均值漂移模型的统计诊断方法,对监测数据的误差数据和强影响数据的统计检验进行分析研究,并以大坝位移监测数据的合理性检验为例进行验证。

1 监测异常数据分析

根据统计诊断中的异常数据分类,结合大坝工程自动化监测数据特点和误差形成原因的不同,将监测数据中的异常值分为随机误差、粗差和系统误差(如图1所示)[2,8]。随机误差主要由各种随机和偶然因素引起,符合均值为零的正态分布,在连续大样本的自动化监测数据中普遍存在,一般不影响正常的统计和时序分析。粗差是指含有粗大误差、严重偏离真实值(或既定统计模型)的数据,常常是由观测过程中的操作疏忽和数据的记录、复制和计算处理过程中的过失错误引起。系统误差是指由相互独立的偶然因素作用引起的监测仪器或监测点故障等所造成的误差,严重偏离真实值(或既定统计模型),常表现为单侧点数据异常波动的现象,并可能具有一定的连续性和阶段性。如观测基点因基础或外力作用产生明显扰动,则会引起观测数据的系统误差。

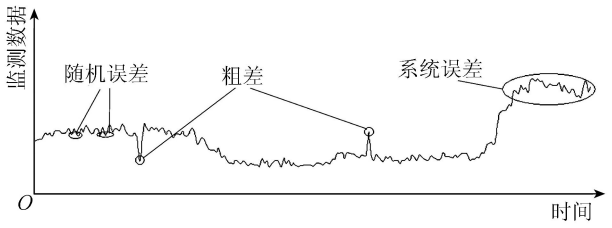

图1 监测异常数据示意图

在数据表现上,粗差具有突发性,在相邻监测数据中通常以个别形式出现,一般不具有连续性;系统误差由于系统故障难以自行修复,往往表现为多个数值接近的测值连续出现,并在均值附近摆动增大,具有一定的趋势性。粗差一般表现为污染正态分布,可采用统计诊断方法进行分析;系统误差往往可通过同类监测数据的综合过程线对比辨识[10],本文不作重点讨论。

在误差分析的基础上,为对监测数据的重要程度进行区分,定义统计诊断中的强影响数据为对统计推断(如统计模型参数、拟合预测值等)影响特别大的监测数据。由于强影响数据对统计诊断结果具有较大影响,需要特别关注。

2 均值漂移模型

大坝监测数据合理性检验的主要目的是删除粗差,并辨识强影响数据。一个很重要的方法就是逐个计算每组数据对回归分析的影响,进而通过考察统计诊断量的方法来获取不同误差的来源。这里采用均值漂移模型对数据进行统计诊断,即在第 i 个数据上增加一个扰动(附加值),这相当于因变量的均值有所漂移,研究这个扰动对估计量及其他统计量影响的显著程度。

对含有自变量 $\boldsymbol{x}_i(\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{i(k-1)}))$ 与因变量 y_i 的 n 次监测资料序列建立线性回归方程:

$$y_i=\boldsymbol{\varphi}_i^{\mathrm{T}}\boldsymbol{\theta}+\varepsilon_i \qquad (i=1,2,\cdots,n) \tag{1}$$

其中 $\boldsymbol{\theta}=(b_0,b_1,\cdots,b_{k-1})^{\mathrm{T}}$ $\boldsymbol{\varphi}_i=(1,\boldsymbol{x}_i)^{\mathrm{T}}$

$$\varepsilon_i\in N(0,\sigma^2)$$

式中:$\boldsymbol{\theta}$ 为回归参数;ε_i 为随机误差项,服从方差为 σ 的标准正态分布;$k-1$ 为自变量 $\boldsymbol{x}_i$ 所包含的元素个数。

记 $\boldsymbol{\Phi}=(\boldsymbol{\varphi}_1^{\mathrm{T}},\boldsymbol{\varphi}_2^{\mathrm{T}},\cdots,\boldsymbol{\varphi}_n^{\mathrm{T}})^{\mathrm{T}}$,$\boldsymbol{Y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$,$\boldsymbol{\varepsilon}=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n)^{\mathrm{T}}$,则式(1)写成矩阵形式如下:

$$\boldsymbol{Y}=\boldsymbol{\Phi\theta}+\boldsymbol{\varepsilon} \tag{2}$$

在最小二乘估计下,未知参数 $\boldsymbol{\theta}$、$\boldsymbol{Y}$、$\boldsymbol{e}$、σ 的最佳估计值 $\hat{\boldsymbol{\theta}}$、$\hat{\boldsymbol{Y}}$、$\hat{\boldsymbol{e}}$、$\hat{\sigma}$ 分别为

$$\hat{\boldsymbol{\theta}}=(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{Y} \tag{3}$$

$$\hat{\boldsymbol{Y}}=\boldsymbol{\Phi}\hat{\boldsymbol{\theta}}=\boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{Y} \tag{4}$$

$$\hat{\boldsymbol{e}}=\boldsymbol{Y}-\hat{\boldsymbol{Y}} \tag{5}$$

$$\hat{\sigma}^2=(n-k)^{-1}\hat{\boldsymbol{e}}^{\mathrm{T}}\hat{\boldsymbol{e}} \tag{6}$$

式中:$\boldsymbol{e}$ 为残差。

记 $\boldsymbol{P}=\boldsymbol{\Phi}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}$(帽子矩阵),其对角元素 p_{ii} 有[11]

$$p_{ii}=\frac{1}{n}+(\tilde{\boldsymbol{\varphi}}_i-\bar{\tilde{\boldsymbol{\varphi}}})^{\mathrm{T}}(\boldsymbol{\Phi}_{\mathrm{c}}^{\mathrm{T}}\boldsymbol{\Phi}_{\mathrm{c}})^{-1}(\tilde{\boldsymbol{\varphi}}_i-\bar{\tilde{\boldsymbol{\varphi}}}) \tag{7}$$

其中 $\boldsymbol{\Phi}_{\mathrm{c}}=\left(I-\frac{1}{n}\boldsymbol{J}\right)\tilde{\boldsymbol{\Phi}}$ $\boldsymbol{J}=\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}$

$$\bar{\tilde{\boldsymbol{\varphi}}}=\frac{1}{n}\sum_{i=1}^{n}\tilde{\boldsymbol{\varphi}}_i \qquad \tilde{\boldsymbol{\varphi}}_i=(x_{i1},x_{i2}\cdots,x_{i(k-1)})^{\mathrm{T}}$$

式中:$\tilde{\boldsymbol{\Phi}}$ 为 $\boldsymbol{\Phi}$ 去掉已知的第1个列向量 $\boldsymbol{1}$ 而得到的矩阵。

式(7)中$(\tilde{\boldsymbol{\varphi}}_i-\bar{\tilde{\boldsymbol{\varphi}}})$称为马氏距离,$\bar{\tilde{\boldsymbol{\varphi}}}_i$ 可看作数据的中心。该式表明,p_{ii} 越大,则第 i 组数据 $\boldsymbol{\varphi}_i$(或 $\tilde{\boldsymbol{\varphi}}_i$)离数据中心越远。这些远离数据中心的点往往就是强影响数据或粗差。

对第 i 个数据 $(y_i,\boldsymbol{\varphi}_i^{\mathrm{T}})$ 增加一个扰动,这时式(1)即为均值漂移模型[8]:

$$\begin{cases} y_j = \boldsymbol{\varphi}_j^{\mathrm{T}}\boldsymbol{\theta} + \varepsilon_j & (j \neq i) \\ y_i = \boldsymbol{\varphi}_i^{\mathrm{T}}\boldsymbol{\theta} + \gamma + \varepsilon_i \end{cases} \tag{8}$$

其矩阵形式为

$$\boldsymbol{Y} = \boldsymbol{\Phi\theta} + \boldsymbol{d}_i\gamma + \boldsymbol{\varepsilon} \tag{9}$$

式中：$\boldsymbol{d}_i$ 为 n 维单位向量，其第 i 个分量为 1，其余均为零；γ 为扰动值。

式(9)相应参数 $\boldsymbol{\theta}$、σ^2、γ 的最小二乘估计分别记为 $\hat{\boldsymbol{\theta}}_{\mathrm{a}}$、$\hat{\sigma}_{\mathrm{a}}^2$、$\hat{\gamma}$，其表达式如下：

$$\hat{\boldsymbol{\theta}}_{\mathrm{a}} = \hat{\boldsymbol{\theta}} - (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\varphi}_i\hat{\gamma} \tag{10}$$

$$\hat{\sigma}_{\mathrm{a}}^2 = \frac{n-k-r_i^2}{n-k-1}\hat{\sigma}^2 \tag{11}$$

$$\hat{\gamma} = \frac{\hat{e}_i}{1-p_{ii}} \tag{12}$$

式中：r_i 为学生化内残差。

均值漂移模型研究的是第 i 组数据 $(y_i, \boldsymbol{\varphi}_i^{\mathrm{T}})$ 对估计量的影响，若 $\hat{\boldsymbol{\theta}}_{\mathrm{a}}$ 与 $\hat{\boldsymbol{\theta}}$ 有显著差异，则说明 $(y_i, \boldsymbol{\varphi}_i^{\mathrm{T}})$ 可能为异常数据(粗差或强影响数据)。由式(10)可以看出，$\hat{\boldsymbol{\theta}}_{\mathrm{a}}$ 与 $\hat{\boldsymbol{\theta}}$ 差值的大小主要取决于扰动估值 $\hat{\gamma}$ 的大小，因此，若其显著异于零，则说明 $(y_i, \boldsymbol{\varphi}_i^{\mathrm{T}})$ 可能异常。

3 基于均值漂移模型的统计诊断

3.1 粗差检验

由式(10)可以看出，第 i 个数据 $(y_i, \boldsymbol{\varphi}_i^{\mathrm{T}})$ 对应的残差 $\hat{e}_i$ 越大，则扰动估值 $\hat{\gamma}$ 越大，估计量 $\hat{\boldsymbol{\theta}}_{\mathrm{a}}$ 与 $\hat{\boldsymbol{\theta}}$ 的差异越大，即第 i 个点对模型的影响越大，因而残差 $\hat{e}_i$ 是决定 $(y_i, \boldsymbol{\varphi}_i^{\mathrm{T}})$ 影响大小很重要的统计量。为消除尺度影响，对残差进行标准化得到其学生化内残差 r_i；基于均值漂移模型，取 $\hat{\sigma}_{\mathrm{a}}^2$ 作为 σ^2 的估计量可得学生化外残差 t_i：

$$\begin{cases} r_i = \dfrac{\hat{e}_i}{\hat{\sigma}\sqrt{1-p_{ii}}} \\ t_i = \dfrac{\hat{e}_i}{\hat{\sigma}_{\mathrm{a}}\sqrt{1-p_{ii}}} \end{cases} \quad (i = 1, \cdots, n) \tag{13}$$

显然，r_i 和 t_i 越大(p_{ii}越大)，$\hat{\boldsymbol{\theta}}_{\mathrm{a}}$ 与 $\hat{\boldsymbol{\theta}}$ 差异越大，这个点的影响也越大。从均值漂移模型出发，若扰动值 γ 显著不等于零，则数据 $(y_i, \boldsymbol{\varphi}_i^{\mathrm{T}})$ 可能存在异常。为此设计如下假设检验：

$$H_0: \gamma = 0; H_1: \gamma \neq 0 \tag{14}$$

若这个假设被否定，则说明 γ 显著异于零，因而模型式(8)成立，即原模型式(1)不成立，可判断 $(y_i, \boldsymbol{\varphi}_i^{\mathrm{T}})$ 为粗差；又由式(10)可知，这时 $\hat{\boldsymbol{\theta}}_{\mathrm{a}}$ 与 $\hat{\boldsymbol{\theta}}$ 之间亦有显著差异，也说明 $(y_i, \boldsymbol{\varphi}_i^{\mathrm{T}})$ 为粗差。

假设检验式(12)的检验函数可由下式给出[12]：

$$F(i) = t_i^2 = \frac{n-k-1}{n-k-r_i^2}r_i^2 = \frac{\hat{e}_i^2}{\hat{\sigma}_{\mathrm{a}}^2(1-p_{ii})} \tag{15}$$

式中：$F(i)$ 服从 F 分布，$F(i) \propto F(1, n-k-1)$；t_i 服从 t 分布，$t_i \propto t(n-k-1)$。给定显著性水平 $1-\alpha$，则检验的势函数为 $\varphi(\gamma) = P_\gamma(F(i) > F(1, n-k-1, 1-\alpha))$ 或 $\varphi(\gamma) = P_\gamma(t_i > t(n-k-1, 1-\alpha))$。即：若 $F(i) > F(1, n-k-1, 1-\alpha)$ 或 $t_i > t_{1-\alpha}(n-k-1)$，则可认为 $(y_i, \boldsymbol{\varphi}_i^{\mathrm{T}})$ 为粗差。

3.2 强影响数据检验

为研究第 i 个数据 $(y_i, \boldsymbol{\varphi}_i^{\mathrm{T}})$ 的影响，仍考虑均值漂移模型式(10)。差值 $(\hat{\boldsymbol{\theta}}_{\mathrm{a}} - \hat{\boldsymbol{\theta}})$ 就是 $(y_i, \boldsymbol{\varphi}_i^{\mathrm{T}})$ 影响大小的一种度量，该差值越大，影响越大。但由于 $(\hat{\boldsymbol{\theta}}_{\mathrm{a}} - \hat{\boldsymbol{\theta}})$ 是一个向量，不便于比较，必须选择一个合适的数量和距离，以便定量比较影响的大小，这里选择应用广泛的 Cook 距离[11,13]。

由线性模型的理论可知，模型式(2)中参数 $\boldsymbol{\theta}$ 的 $1-\alpha$ 置信域可表示为一个椭球的形式：

$$\frac{(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}{k\hat{\sigma}^2} \leqslant F(k, n-k, 1-\alpha) \tag{16}$$

式中：$F(k, n-k, 1-\alpha)$ 表示 F 分布的 $1-\alpha$ 分位点。

在参数空间 $\boldsymbol{R}^k$ 中，式(16)表示一个以 $\hat{\boldsymbol{\theta}}$ 为中心的椭球。易见，落在椭球以外的 $\boldsymbol{\theta}$ 点可能性很小，其概率只有 α。现考虑 $\hat{\boldsymbol{\theta}}_{\mathrm{a}}$，如果落在椭球之外，则说明 $\hat{\boldsymbol{\theta}}_{\mathrm{a}}$ 与 $\hat{\boldsymbol{\theta}}$ 差异非常大，将 $(y_i, \boldsymbol{\varphi}_i^{\mathrm{T}})$ 作为模型式(2)的数据是不可接受的。同理，若将 $\hat{\boldsymbol{\theta}}_{\mathrm{a}}$ 的值代入式(16)左端所得的值较大，则说明 $\hat{\boldsymbol{\theta}}_{\mathrm{a}}$ 离置信域中心 $\hat{\boldsymbol{\theta}}$ 较远，因而 $\hat{\boldsymbol{\theta}}_{\mathrm{a}}$ 与 $\hat{\boldsymbol{\theta}}$ 有较大差异，从而可认为 $(y_i, \boldsymbol{\varphi}_i^{\mathrm{T}})$ 对模型式(2)的影响也较大。基于这种考虑，Cook 于 1977 年提出在式(16)左端以 $\hat{\boldsymbol{\theta}}_{\mathrm{a}}$ 代替 $\hat{\boldsymbol{\theta}}$，作为度量第 i 个数据点 $(y_i, \boldsymbol{\varphi}_i^{\mathrm{T}})$ 影响大小的数量指标。由此，将第 i 个数据点 $(y_i, \boldsymbol{\varphi}_i^{\mathrm{T}})$ 的 Cook 距离 D_i 定义为

$$D_i = \frac{(\hat{\boldsymbol{\theta}}_{\mathrm{a}} - \hat{\boldsymbol{\theta}})^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}(\hat{\boldsymbol{\theta}}_{\mathrm{a}} - \hat{\boldsymbol{\theta}})}{k\hat{\sigma}^2} \tag{17}$$

由式(17)可见，D_i 表示 $\hat{\boldsymbol{\theta}}_{\mathrm{a}}$ 与 $\hat{\boldsymbol{\theta}}$ 的一种加权距离，其权因子为 $\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}/(k\hat{\sigma}^2)$。另外，由于式(17)分母上有 $k\hat{\sigma}^2$，因而 D_i 与尺度无关。

将式(10)和式(12)代入式(17)中，D_i 又可表示成如下形式：

$$D_i = \frac{p_{ii}}{1-p_{ii}}\frac{r_i^2}{k} \tag{18}$$

式(18)和式(13)说明，Cook 距离 D_i 的大小取决于残差 $\hat{e}_i$ 以及帽子矩阵对角元素 p_{ii}的大小，前者反映拟合情况，后者表示第 i 个数据距离数据中心的远近程度。从式(18)也可以看出，p_{ii}大的数据不一定 D_i 大，这也说明粗差不一定就是强影响数据。

4 实例分析

选取某重力坝坝顶引张线测点顺河向位移自动化监测数据为例，采用上述统计诊断方法进行监测数据的合理性检验。分析选用的典型测值过程线如图2所示，时间序列为2006年12月22日至2012年9月20日，测值以向下游为“+”，向上游为“-”。

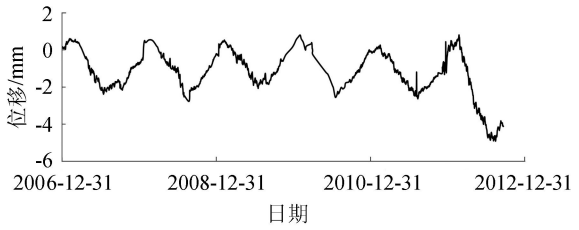

图2 坝顶测点顺河向位移过程线

根据测值过程线可以看出，坝顶顺河向位移呈较为明显的年周期变化，受水位、温度影响显著，考虑时效因素影响，其位移监测资料的统计模型可表征为如下形式[14]：

$$y = y_H + y_T + y_\theta = \sum_{i=1}^{3}[a_i(H_u - H_{u0})] + \sum_{i=1}^{2}\left[b_{1i}\left(\sin\frac{2\pi it}{365} - \sin\frac{2\pi it_0}{365}\right) + b_{2i}\left(\cos\frac{2\pi it}{365} - \cos\frac{2\pi it_0}{365}\right)\right] + c_1(t - t_0) + c_2(\ln t - \ln t_0) + a_0 + \varepsilon \quad (19)$$

其中 $\varepsilon \in N(0,\sigma^2)$

式中：y_H, y_T, y_θ 分别为水压分量、温度分量、时效分量；H_u, H_{u0} 分别为监测日、始测日所对应的上游水头；a_i 为水压因子回归系数；t 为位移监测日至始测日的累计天数；t_0 为建模资料系列第一个监测日至始测日的累计天数；b_{1i}, b_{2i} 为温度因子回归系数；c_1，c_2 为时效因子回归系数；a_0 为常数项。

比照式(1)，监测数据向量为

$$\boldsymbol{\varphi}_i = \left(1, H_u - H_{u0}, H_u^2 - H_{u0}^2, H_u^3 - H_{u0}^3, \sin\frac{2\pi t}{365} - \sin\frac{2\pi t_0}{365}, \cos\frac{2\pi t}{365} - \cos\frac{2\pi t_0}{365}, \sin\frac{4\pi t}{365} - \sin\frac{4\pi t_0}{365}, \cos\frac{4\pi t}{365} - \cos\frac{4\pi t_0}{365}, t - t_0, \ln t - \ln t_0\right)^{\mathrm{T}} \quad (20)$$

未知参数为

$$\boldsymbol{\theta} = (a_0, a_1, a_2, a_3, b_{11}, b_{21}, b_{12}, b_{22}, c_1, c_2)^{\mathrm{T}} \quad (21)$$

监测数据组数 $n=1\,634$（部分时段无测值），$k=10$。

根据测值变化规律，2012年第二季度开始，测值减少较为显著。综合同类测点监测资料及同期环境量变化分析，认为测值显著变化是由右岸观测基点的位移造成，属系统误差数据，一并对其进行计算分析。

对所有监测数据采用本文检验方法进行统计诊断，部分异常数据检验结果如表1所示。

表1 部分数据异常情况的检验结果

序号	日期	y/mm	$\hat{y}$/mm	$\hat{e}_i$/mm	t_i	Cook距离
1	2008-01-08	-0.681	0.383	-1.064	-2.081	0.003 0
2	2008-01-16	-0.537	0.519	-1.056	-2.065	0.003 0
3	2009-01-11	-0.325	0.526	-0.851	-1.664	0.000 8
4	2009-04-06	-0.670	0.170	-0.840	-1.642	0.000 4
5	2009-10-31	-0.544	-1.329	0.785	1.536	0.000 8
6	2010-07-08	-2.040	-2.717	0.677	1.337	0.004 0
7	2011-08-05	-1.194	-2.766	1.572	3.080	0.006 0
8	2011-12-20	0.440	-1.015	1.455	2.849	0.007 0

如取 $\alpha=0.05$，由检验函数可得 $t_{0.95}(1\,634)=1.679$。由此对数据检验结果进行评判，将 t_i 绝对值大于上述临界值的数据判断为粗差。对第一次检验粗差剔除后的数据再重新建模检验，直到剩余数据满足t检验(2012年系统误差数据暂不处理)，由此共删除粗差13个，删除率0.8%。

按照式(3)~(6)对模型拟合效果进行计算分析，实测值、拟合值过程线如图3所示。删除粗差后，模型拟合精度与原始数据精度相比有所提高，复相关系数 R 从0.904提高到0.912，剩余标准差 S 从0.512降低到0.468，说明了统计诊断识别粗差的有效性。

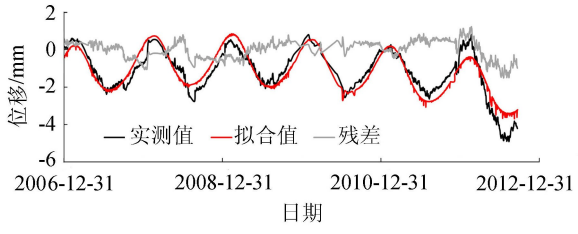

图3 测点实测值、拟合值及残差过程线

删除粗差后各测值的Cook距离如图4所示。可以看出2012年后的Cook距离计算值较大，与该时段存在系统误差数据的原因相符。监测资料的初始阶段Cook距离计算值也较大，说明初始阶段测值对建模的影响较大，应尽量减少该时段的观测误差。而在运行期，Cook距离计算值较大区域一般出现在每年的七八月份，该时期受强降雨影响，水库水位变

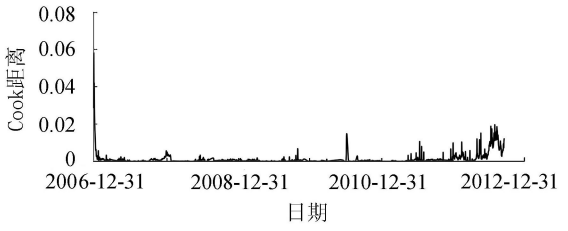

图4 各测值Cook距离计算值过程线

动较大,计算值较好地反映了环境变化对大坝位移的影响。

5 结 语

a. 对大坝安全监测的异常数据分类进行分析,结合误差数据形成原因的不同,划分为随机误差、粗差、系统误差等,并辨识强影响数据。

b. 基于统计诊断的均值漂移模型,研究了不同异常数据的处理方法,包括以模型扰动值为依据的粗差的 t 检验法和以模型扰动对拟合参数的影响为依据的强影响数据的 Cook 距离检验法。

c. 以典型大坝的位移自动化监测数据为例,采用本文统计诊断方法对监测数据进行了合理性检验,结果表明该方法可有效辨识粗差和强影响数据,能提高数据建模拟合的精度和进一步分析的准确性。

参考文献:

[1] 顾冲时, 苏怀智. 混凝土坝工程长效服役与风险评定研究述评[J]. 水利水电科技进展, 2015,35(5):1-12. (GU Chongshi, SU Huaizhi. Current status and prospects of long-term service and risk assessment of concrete dams [J]. Advances in Science and Technology of Water Resources, 2015, 35(5):1-12. (in Chinese))

[2] 李子阳, 马福恒, 华伟南. 多源信息融合诊断大坝安全监测资料合理性[J]. 水利水运工程学报, 2013(1):41-46. (LI Ziyang, MA Fuheng, HUA Weinan. The diagnosis of dam safety monitoring data rationality based on multiple source information fusion [J]. Hydro-science and Engineering, 2013(1):41-46. (in Chinese))

[3] HRISTOPULOS D T, MERTIKAS S P, ARHONTAKIS I, et al. Using GPS for monitoring tall-building response to wind loading: filtering of abrupt changes and low-frequency noise, variography and spectral analysis of displacements[J]. GPS Solut, 2007(11):85-95.

[4] SHARP I, YU K. Indoor TOA error measurement, modeling, and analysis [J]. IEEE Transactions on Instrumentation and Measurement, 2014, 63(9):2129-2144.

[5] JIANG Xiaolong, LIU Pei, LI Zheng. Data reconciliation and gross error detection for operational data in power plants[J]. Energy, 2014,75:14-23.

[6] TAYLOR J R, LOESCHER H L. Automated quality control methods for sensor data: a novel observatory approach [J]. Biogeosciences Discussions, 2013, 10: 4957-4971.

[7] 胡江, 苏怀智, 马福恒,等. MF-DFA 在大坝安全监测序列分析和整体性态识别中的应用[J]. 水利水电科技进展, 2014, 34(3):50-55. (HU Jiang, SU Huaizhi, MA Fuheng, et al. The application of MF-DFA in time series analysis and global state recognition of dam safety monitoring[J]. Advances in Science and Technology of Water Resources, 2014, 34(3):50-55. (in Chinese))

[8] 韦博成, 林金官, 解锋昌. 统计诊断[M]. 北京:高等教育出版社, 2009.

[9] 郭丽, 袁永生, 李子阳. 大坝滑坡体监测数据的统计诊断[J]. 水电自动化与大坝监测, 2007, 31(4):51-53. (GUO Li, YUAN Yongsheng, LI Ziyang. Statistical diagnosis for the monitoring data of dam slipmass[J]. Hydropower Automation and Dam Monitoring, 2007, 31(4):51-53. (in Chinese))

[10] 胡波, 刘观标, 吴中如. 小湾特高拱坝首蓄期坝体变形特性分析及评价[J]. 水利水电科技进展, 2015, 35(6):68-72. (HU Bo, LIU Guanbiao, WU Zhongru. Deformation analysis and safety evaluation of Xiaowan Ultra-high Arch Dam during first impoundment [J]. Advances in Science and Technology of Water Resources, 2015, 35(6):68-72. (in Chinese))

[11] COOK R D, WEISBERG S. Residuals and Influence in Regression[M]. New York:Chapman and Hall,1982.

[12] FERGUSON T S. On the rejection of outliers [J]. Proceedings of the Fourth Berkley Symposium in Mathematical Statistics and Probability,1961,2(2): 123-146.

[13] COOK R D. Influential observation in linear regression [J]. Journal of the American Statistical Association, 1979, 74:169-174.

[14] 吴中如. 水工建筑物安全监控理论及其应用[M]. 北京: 高等教育出版社, 2003.

(收稿日期:2018-04-28 编辑:骆超)

(上接第 58 页)

[22] CORINNA C,VLADIMIR V. Support-vetor networks[J]. Machnine Learning, 1995, 20:273-297.

[23] 马春辉, 杨杰, 程琳, 等. 基于混合核函数 HS-RVM 的边坡稳定性分析[J]. 岩石力学与工程学报, 2017, 36(增刊 1): 3409-3415. (MA Chunhui, YANG Jie, CHENG Lin, et al. Slope stability analysis based on HS-RVM with mixed kernel[J]. Chinese Journal of Rock Mechanics and Engineering, 2017, 36(Sup1): 3409-3415. (in Chinese))

[24] 蒋水华, 李典庆, 周创兵. 基于拉丁超立方抽样的边坡可靠度分析非侵入式随机有限元法[J]. 岩土工程学报, 2013, 35(增刊 2):70-76. (JIANG Shuihua, LI Dianqing, ZHOU Chuangbing. Non-intrusive stochastic finite element method for slope reliability analysis based on Latin hypercube sampling[J]. Chinese Journal of Geotechnical Engineering, 2013, 35(Sup2): 70-76. (in Chinese))

[25] 邵磊. 基于裂缝扩展细观模拟的堆石料流变特性研究[D]. 大连:大连理工大学,2013.

(收稿日期:2018-05-03 编辑:雷燕)