DOI:10. 19431/j. cnki. 1673-0062. 2021. 06. 006

尾矿坝振动实验数据样本容量的优化研究

刘振广, 徐正华, 廖茂新*, 刘宏亮

(南华大学 数理学院,湖南 衡阳 421001)

摘　要:针对铀尾矿坝振动实验逸出氡累积浓度数据样本容量设计的主观性,本文基于偏向角绝对值均值递减还原逸出氡累积浓度曲线,采用随机起点等距抽样方法,得到不同样本容量的数据子集;通过曲线拟合和样本决定系数度量不同子集对还原实验数据集的代表能力。在保障实验研究精度条件下,得到使实验成本较优的最小样本容量,实验结果表明在样本决定系数大于 0.98 时,优化后样本容量为实验数据样本容量的 10% 以下。

关键词:逸出氡累积浓度;随机起点;等距抽样;样本决定系数;样本容量
中图分类号:TD868　　**文献标志码**:A
文章编号:1673-0062(2021)06-0035-05

Research on Optimization of Sample Size of Tailing Dam Vibration Experiment

LIU Zhenguang, XU Zhenghua, LIAO Maoxin*, LIU Hongliang
(School of Mathematics and Physics, University of South China, Hengyang, Hunan 421001, China)

Abstract: To avoid the subjectivity of sample size design in the vibration experiment of uranium tailing dam, the method which describes the mean of absolute value of deviation angle is applied to restore the cumulative concentration curve of escaping radon. Sample subsets of varied capacity are obtained by isometric sampling with random starting point. The coefficient of determination of curve fitting is used to evaluate the representative ability to restore the experimental data sets. Without detriment to the accuracy of the experiment, the minimum sample size that cost less is obtained. The result shows that the subset capacity is 10% less that of the whole data sets when the determination coefficient is over 0.98.

key words: cumulative concentration of escaping radon; random starting point; isometric

收稿日期:2021-07-03
基金项目:湖南省研究生科研创新项目(CX20200925);湖南省教育厅重点项目(20A440;20A425);国防科工局关于技术基础科研"十三五"第六批项目(403C001);湖南省环保科研课题(湘财建二指(2019)0011号);湖南省研究生教改项目(2019JGYB192)
作者简介:刘振广(1992—),男,硕士研究生,主要从事数据处理方面的研究。E-mail:598494081@qq.com。*通信作者:廖茂新(1969—),男,教授,博士,主要从事微分方程方面的研究。E-mail:841139745@qq.com

sampling method;determination coefficient of sample;sample size

0 引 言

尾矿是矿石经矿场筛选后剩余的砂状废弃物,尾矿逐渐堆积形成尾矿坝[1]。振动造成的尾矿坝结构变化是造成溃坝事故的重要原因之一。国内外对振动条件下尾矿坝稳定性进行了多方向的探索,有研究地震条件下尾矿坝事故的,有通过室内振动实验研究尾矿坝模型变化的,其中,KULALI[2]通过实验分析了不同土壤对氡析出的影响。蔡嗣经、张栋等[3]通过实验室振动实验研究了尾矿砂的动力特性。柳厚祥、廖雪等[4]研究了地震对尾矿坝裂缝水压变化的影响。陈存礼、何军芳[5]等研究了三轴实验中饱和尾矿砂动孔压变化规律。A. D. K. Tareen[6]等人通过运用箱线图对氡逸出率实时数据异常变化的检测,实现了基于氡析出率变化的地震预测方法。C. Bilibio[7]与 S. Cockenpot[8]研究了温度对氡析出率的影响,得到氡析出率随温度升高而增加的规律。马宇艇[1]通过振动台系统研究了尾矿坝振动动力响应,得到尾矿坝变形破坏的基本趋势。

本文根据振动实验设计的需要,研究数据样本容量的优化度量方法。在实验数据总体统计特征已知的条件下样本容量可以根据统计学公式来确定[9];在实验数据特征未知的条件下,样本容量通常根据以往经验或相似实验来确定。振动实验中逸出氡累积浓度数据的特征具有未知性,按相似实验经验确定样本容量将造成实验数据失真或实验成本过高等问题。

本文在两种振动实验条件上,对实验数据集进行数据还原后抽样获得数据子集,对数据子集进行曲线拟合得到拟合方程,度量拟合方程对还原后数据集的代表能力,在满足实验精度要求的前提下得到实验较优优样本容量。

1 实验介绍与样本容量优化度量算法设计

1.1 实验数据采集与还原

尾矿坝振动模拟实验主要包括测氡仪、双向激振系统、逸出集氡模型试验箱、信号采集处理系统四部分组成。其中测氡仪用来测量氡浓度,激振系统用来对模型施加激振力,逸出氡模型实验箱用来收集逸出气氡,信号采集处理系统用来接收并显示振动与氡密度信号。在设定电压、电流及振动频率的条件下,通过对气氡浓度数据进行等距采样,设定采样间距为 5 min,得到实验数据样本集 $B=\{(t_i,y_i)\mid i=1,2,\cdots,36\}$,其中,$t_i$ 表示时间点,y_i 表示时间点对应的逸出氡累积浓度。

振动条件下微塑模型中,由于尾砂量有限,累积逸出氡累积浓度变化遵循气体扩散效应,初始阶段累积氡浓度大致呈线性趋势匀速增长,然后累积氡浓度增长速度递减而渐趋平稳。

实验条件的限制导致振动实验时间不够长,采集到的样本数据量较小,数据基本分布在逸出效应曲线的中前段。本文首先在实验样本数据与逸出效应规律的基础上,通过函数构造法对实际数据进行反演还原。本文选取偏向角均值递减方法对两种振动条件下的实验数据集进行还原,得到振动条件下累积氡浓度数据集 $D=\{(t_i,y_i)\mid i=1,2,\cdots,n\}$。

本文构造了偏向角均值递减的数据还原方法,利用数据集 B 中数据偏向角的变化规律,通过计算偏向角均值并使偏向角按均值递减,推导还原出实验未能得到的逸出氡累积浓度值。

设

$$\theta_i=\frac{\arctan(y_{i+1}-y_i)}{t_{i+1}-t_i},i=1,2,\cdots,35, \quad (1)$$

即 θ_i 表示数据集 B 中的倾斜角[10]。

令

$$\Delta\bar{\theta}=(\theta_{35}-\theta_1)/35, \quad (2)$$

即 $\Delta\bar{\theta}$ 表示偏向角均值。

设还原数据偏向角按照均值递减,可得

$$\theta_i=\theta_{i-1}-\Delta\bar{\theta},i=36,37,\cdots,n-1, \quad (3)$$

由偏向角均值递减可推导得到

$$y_i=y_{i-1}+(t_i-t_{i-1})\tan\theta_{i-1},i=37,38,\cdots,n。 \quad (4)$$

对还原后实验数据集绘制散点图,如图 1 和图 2 所示,图 1 中电压为 3 V、电流为 1 A、振动频率为 10 Hz 的条件下逸出氡累积浓度随时间变化趋势散点图;图 2 中电压为 3 V、电流为 2 A、振动频率为 30 Hz 的条件下逸出氡累积浓度随时间变化趋势散点图。

1.2 确定最小样本容量

本文通过对还原后实验数据集 D 进行不同间隔距离下的等距抽样,得到不同样本容量的数据子集 S。根据还原后实验数据散点图的特征,

构造了反正切型函数来模拟逸出氡累积浓度曲线,利用最小二乘拟合思想得到样本子集 S 的拟合方程。通过样本判定系数 R_C^2 度量拟合方程对子集 S 的拟合优度,得到样本子集 S 对数据集 D 的代表能力,在保证实验精度要求的情况下得到较优样本容量。

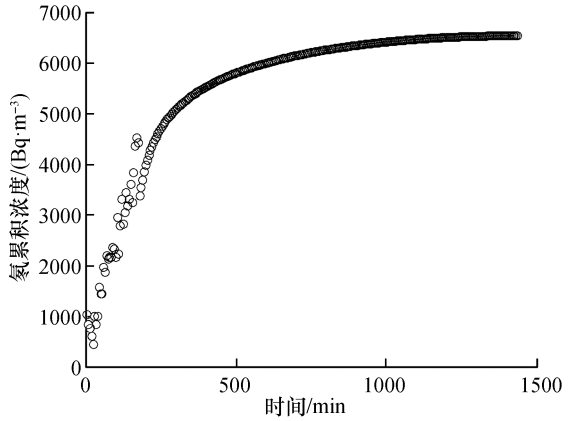

图 1　振动条件为 3 V-1 A-10 Hz 时氡累积浓度变化散点图

Fig. 1　Scatter plot of the cumulative density of radon when the vibration is 3 V-1 A-10 Hz

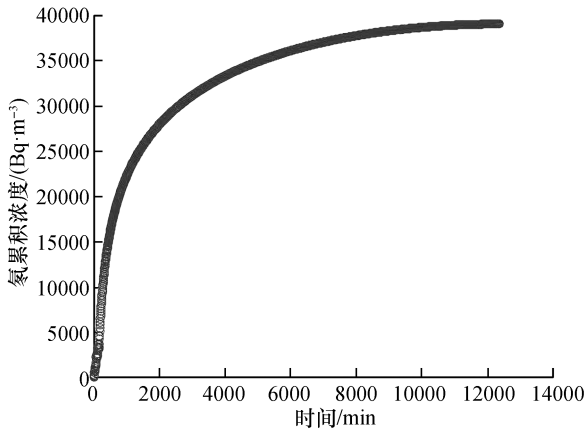

图 2　振动条件为 3 V-2 A-30 Hz 氡累积浓度变化散点图

Fig. 2　Scatter plot of the cumulative density of radon when the vibration is 3 V-2 A-30 Hz

1.2.1　随机起点等距抽样

等距抽样又称为系统抽样,是将总体平均按一定顺序分为若干部分,随机确定起点后分别从各部分中抽取数据形成样本子集 S。等距抽样适用于总体数据具有一定稳定变化趋势的情况,可以保证所抽取的样本代表性较好[11]。本文应用随机起点等距抽样方法对还原后实验数据集 D 进行抽样,在不同采样间距下得到样本子集 $S=\{(x_j,c_j)\mid j=1,2,\cdots,m\}$。

设 l 为采样间距,其中 $l=5,10,\cdots,5k,k\in Z$,则样本容量 m 为

$$m=\begin{cases}\left[\dfrac{n}{l}\right] & 1\leqslant i_1\leqslant n\%l\\ \left[\dfrac{n}{l}\right] & n\%l<i_1\leqslant\dfrac{l}{5}\end{cases}\tag{5}$$

式中 n 是数据集 D 的元素数;$n\%l$ 表示 n 除以 l 的余数;$\left[\dfrac{n}{l}\right]$ 表示 n 除以 l 向上取整数;$\left[\dfrac{n}{l}\right]$ 表示 n 除以 l 向下取整数。随机等距抽样算法 1.1 如下:

算法 1.1　随机起点等距抽样算法

步骤 1　输入还原后数据集 $D=\{(t_i,y_i)\mid i=1,2,\cdots,n\}$ 和采样间距 l,($l=5,10,\cdots,60$);

步骤 2　从数据集 D 前$\dfrac{l}{5}$元素中随机抽取一个,设被抽取元素编号为 i_1,$1\leqslant i_1\leqslant\dfrac{l}{5}$,则样本子集 S 中第一个元素$(x_1,c_1)=(t_{i_1},c_{i_1})$,$1\leqslant i_1\leqslant\dfrac{l}{5}$;

步骤 3　顺序抽出数据集 D 编号为 $i_1,i_1+l/5,\cdots,i_1+(m-1)l/5$ 的元素,则样本子集 S 中元素$(x_j,c_j)=(t_{i_1+(j-1)l/5},c_{i_1+(j-1)l/5})$,$1\leqslant i_1\leqslant l/5$,$j=1,2,\cdots,m$;

步骤 4　输出样本子集 $S=\{(x_j,c_j)\mid i=1,2,\cdots,m\}$。

在不同采样间距下抽取出不同样本容量的数据样本,下面给出不同样本最小二乘拟合与拟合效果评价的方法。

1.2.2　反正切型函数最小二乘估计

在还原后数据集图像中,图像先线性上升然后上升速度渐缓最终比较平稳,实际数据分布如图 1 和图 2 所示。

由图 1 和图 2 图像可以观察到数据集 D 与反正切型函数图像相似,因此本文采用反正切型函数模型对抽取出的样本子集 S 进行拟合,得到不同样本子集对应的拟合方程。

设反正切型函数回归方程为

$$y=a+b\times a\tan(c\times x),\tag{6}$$

式中 a,b,c 为待定参数,待定参数需要通过样本数据估计得到。

对于待定参数的估计,本文采用最小二乘法,待定参数使得模型估计值 $\hat{y}_j$ 与数据实际值 y_j 之

间残差的平方和最小,即得到拟合数据相对较优的曲线[12]。由最小二乘法可得到

$$Q = \sum_{j=1}^{m} (y_j - \hat{y}_j)^2 = \sum_{j=1}^{m} (y_j - (a + b \times a\tan(c \times x)))^2, \quad (7)$$

式中 Q 最小时,对应待定参数即为较优拟合曲线的最小二乘参数 $\hat{a},\hat{b},\hat{c}$。

1.2.3 修正的样本决定系数 R_C^2

曲线拟合优度通常使用决定系数 R^2 来表示[13],即

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (8)$$

式中 SST 中由于 $\sum_{i=1}^{n} (y_i - \bar{y})^2$ 表示实际值与均值的平方和。在曲线情况下由于均值与实际值的偏差较大,且平方放大了这种偏差,导致 R^2 对于曲线拟合优度的灵敏度过高。因此本文根据实验数据的实际需求,设计了一种度量曲线拟合优度的指标即样本决定系数 R_C^2。

本文首先对样本子集 S 进行拟合得到反正切型函数曲线 $y=\hat{a}+\hat{b}\times a\tan(\hat{c}\times x)$,通过拟合曲线 $y=\hat{a}+\hat{b}\times a\tan(\hat{c}\times x)$ 对数据集 D 的拟合优度表示样本子集 S 对总体数据集的代表能力。曲线拟合优度指标为样本决定系数 $R_C^2, R_C^2 \in (-\infty, 1)$。$R_C^2$ 公式如下,

$$R_C^2 = 1 - \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} y_i}, \quad (9)$$

式中 y_i 和 $\hat{y}_i$ 分别是时间点 $t_i, (i=1,2,\cdots,n)$ 对应的逸出氡累积浓度观测值与估计值。

当 y_i 和 $\hat{y}_i$ 接近程度很高时,则离差绝对值之和 $\sum_{i=1}^{n} |y_i - \hat{y}_i|$ 越小,所以 $\sum_{i=1}^{n} |y_i - \hat{y}_i|$ 与逸出氡累积浓度观测值之和 $\sum_{i=1}^{n} y_i$ 的比值也越小,故 R_C^2 越大。当 y_i 和 $\hat{y}_i$ 完全一致时,则 R_C^2 取最大值为1,表示拟合曲线 $y = \hat{a} + \hat{b} \times a\tan(\hat{c} \times x)$ 对数据集 D 的拟合优度达到最高。

本实验样本决定系数 $R_C^2 \geqslant 0.98$。

1.3 样本容量优化度量算法设计

已知还原后逸出氡累积浓度数据集 $D=\{(t_i, y_i) \mid i=1,2,\cdots,n\}$,得到散点图匹配对应的拟合方程。设定不同的采样间距 l,从数据集 D 中抽取样本子集,通过对样本子集数据函数计算样本决定系数,得到在保证精度 $R_C^2 \geqslant 0.98$ 条件下的较优样本容量。详细过程见样本容量优化度量算法 1.2。

算法 1.2 样本容量优化度量算法

步骤 1 输入还原后数据集 $D=\{(t_i, y_i) \mid i=1,2,\cdots,n\}$;

步骤 2 绘制数据集 D 散点图,匹配反正切型函数拟合模型 $y=a+b\times a\tan(c\times x)$;

步骤 3 从数据集 D 中采集元素组成数据样本子集 S;

步骤 3.1 调用算法 1.1;

步骤 3.2 输出样本子集 $S=\{(x_j, c_j) \mid j=1,2,\cdots,m\}$。

步骤 4 对样本子集 S 进行反正切型函数最小二乘拟合得到拟合参数 $\hat{a},\hat{b},\hat{c}$ 与拟合方程 $y=\hat{a}+\hat{b}\times a\tan(\hat{c}\times x)$;

步骤 5 将数据集 D 代入拟合方程 $y=\hat{a}+\hat{b}\times a\tan(\hat{c}\times x)$,计算 R_C^2;

步骤 6 输出样本决定系数 R_C^2。

2 数据分析

对还原后数据集 D 进行抽样得到数据子集 S。对数据子集 S 进行反正切型函数拟合得到拟合方程 $y=\hat{a}+\hat{b}\times a\tan(\hat{c}\times x)$。将数据集 D 代入拟合方程 $y=\hat{a}+\hat{b}\times a\tan(\hat{c}\times x)$ 并计算曲线拟合优度指标 R_C^2。

表 1 和表 2 分别是不同振动条件下还原后数据集及其抽样子集的反正切型函数样本判定系数 R_C^2 结果。

表 1 振动条件为 3 V-1 A-10 Hz 时不同样本容量的 R_C^2

Table 1 R_C^2 for different sample sizes when the vibration is 3 V-1 A-10 Hz

样本容量	287	144	95	72	48	36	24
抽样间距/min	5	10	15	20	30	40	60
R_C^2	0.991 20	0.988 01	0.988 68	0.988 75	0.986 37	0.986 61	0.987 55

表 2　振动条件为 3 V-2A-30 Hz 时不同样本容量的样本判定系数

Table 2　R_C^2 for different sample sizes when the vibration is 3 V-2 A-30 Hz

样本容量	抽样间距	R_C^2	样本容量	抽样间距	R_C^2
2 473	5	0.984 60	247	50	0.980 06
1 236	10	0.979 91	225	55	0.979 63
825	15	0.979 92	190	65	0.979 71
618	20	0.980 02	165	75	0.979 30
494	25	0.980 13	145	80	0.980 08
412	30	0.980 13	130	95	0.980 60
353	35	0.980 14	112	110	0.981 06
309	40	0.979 78	91	135	0.981 20
275	45	0.979 63	54	230	0.979 12

3　结果与讨论

1)在 3 V-1 A-10 Hz 振动条件下,数据集 D 采样间距为 5 min,样本容量为 287,决定系数 R_C^2 为 0.984 60,均方误差 MSE 为 17 253.29。在 $R_C^2 \geqslant 0.98$ 标准下取样本容量最小的数据子集即为较优样本容量子集,其采样间距为 60 min,较优样本容量为 24,决定系数 R_C^2 为 0.987 55,均方误差 MSE 为 27 496.85。

2)在 3 V-2 A-30 Hz 振动条件下,数据集 D 采样间距为 5 min,样本容量为 2 473,决定系数 R_C^2 为 0.984 60,均方误差 MSE 为 817 710.32。在 $R_C^2 \geqslant 0.98$ 标准下取样本容量最小的数据子集即为较优样本容量子集,其采样间距为 135 min,较优样本容量为 91,决定系数 R_C^2 为 0.981 20,均方误差 MSE 为 824 673.97。

参考文献:

[1] 马宇艇. 振动对尾矿坝安全影响的室内试验研究[D]. 北京:首都经济贸易大学,2013:2-4.

[2] KULALI F, AKKURT I. The effect of meteorological parameters on radon concentration in soil gas[J]. Acta physica polonica, 2017, 132(3): 999-1001.

[3] 蔡嗣经,张栋,何理. 地震中尾矿库液化失稳机理及数值模拟研究[J]. 有色金属科学与工程,2011,2(2):1-6.

[4] 柳厚祥. 高尾矿坝的有效应力地震反应分析[J]. 振动与冲击,2008,27(1):65-92.

[5] 陈存礼,何军芳,胡再强,等. 动荷作用下饱和尾矿砂孔压和残余应变演化特性[J]. 岩石力学与工程学报,2006,25(增刊2):4034-4039.

[6] TAREEN A D K, NADEEM M S A, KEARFOTT K J, et al. Descriptive analysis and earthquake prediction using boxplot interpretation of soil radon time series data[J]. Appliedradiation and isotopes, 2019, 154: 108861.

[7] BILIBIO C, SCHELLERT C, RETZ S, et al. Water balance assessment of different substrates on potash tailings piles using nonweighable lysimeters[J]. Journal of environmental management, 2017, 196: 633-643.

[8] COCKENPOT S, CLAUDE C, RADAKOVITCH O. Estimation of air-water gas exchange coefficient in a shallow lagoon based on ^{222}Rn mass balance[J]. Journal of environmental radioactivity, 2015, 143: 58-69.

[9] 邵志强. 抽样调查中样本容量的确定方法[J]. 统计与决策,2012(22):12-14.

[10] 宗凯,符世琛,李一鸣,等. 截割头载荷对掘进机机身偏向角的影响规律分析[J]. 工矿自动化,2018,44(8):46-51.

[11] 金勇进. 抽样调查[M]. 北京:高等教育出版社,2015:99.

[12] 贾俊平. 统计学[M]. 北京:中国人民大学出版社,2018:183-184.

[13] 谢兰,高东红. 非线性回归方法的应用与比较[J]. 数学的实践与认识,2009,39(10):117-121.