不确定海洋数据的质量抽样检验模型研究

王振华[1,2] 周雪楠[1,2] 黄冬梅[1]
(上海海洋大学信息学院 上海 201306)[1]
(海洋赤潮灾害立体监测技术与应用国家海洋局重点实验室 上海 200135)[2]

摘 要 海洋数据具有海量、多源、多类和多维等特性,其数据质量具有不确定性现象,传统的抽样检验理论不能满足海洋数据质量检验的需求。在抽样检验模型的制定中引入了梯形模糊数的思想,将抽样检验模型的制定抽象为模糊线性规划问题,建立了模糊的质量抽样检验模型,解决了具有不确定质量参数海洋数据的质量检验问题,从而完善了抽样检验模型的系统体系。

关键词 梯形模糊数,质量控制,抽样检验模型
中图法分类号 TP274+.3 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.2.039

Sampling Model for Quality Inspection of Uncertain Ocean Data

WANG Zhen-hua[1,2] ZHOU Xue-nan[1,2] HUANG Dong-mei[1]
(College of Information Technology, Shanghai Ocean University, Shanghai 201306, China)[1]
(Key Laboratory of Integrated Marine Monitoring and Applied Technologies for Harmful Algal Blooms, S. O. A., MATHAB, Shanghai 200135, China)[2]

Abstract Ocean data are characterized by magnanimity, multisource, various types and multi-dimensional, so mostly the quality characteristic of ocean data is uncertain. Thus, the conventional theory of sampling inspection cannot satisfy the requirement of quality inspection for ocean data. In this paper, a fuzzy sampling model was proposed based on trapezoid fuzzy number. The fuzzy sampling model has advantage of the performing quality inspection for ocean data, which have uncertain quality characters, and improves the conventional theory of sampling inspection.

Keywords Trapezoid fuzzy number, Quality control, Sampling model

1 引言

随着观测技术的发展,海洋数据的获取手段出现了多样化,包括:空中监测平台,地面、海面监测平台以及海底监测平台(水下传感器)等。因此,海洋数据从GB、TB到PB量级急速增长,伴随而生的海量海洋质量控制问题成为研究热点。

抽样检验是实施质量控制的重要手段之一[1],其原理是以"用尽量少的样本量来尽量准确地评判总体(批)",使检验费用和检验精度达到一种平衡[2]。美国学者Dodge和Roming是现代抽样检验理论的创始人,其给定了检验产品的接收质量限(AOQ),并推导了一系列优化的抽样检验模型[3-5]。在此基础上,很多学者基于概率和数理统计理论,结合待检验批的不合格品率,提出和设计了抽样检验模型,文献[6]通过采用100%检验,获取了检验批的先验知识,然后基于马尔可夫链推导了二阶抽样检验模型。文献[7]利用非线性规划理论,通过抽检特性函数曲线(OC曲线)上的控制点推导了优化的抽样检验模型。文献[8]通过控制检验费用设计了连续抽样检验模型,该模型根据产品的质量特性随时调整和优化抽样模型。文献[9]基于分层抽样的思想,详细阐述了一种"Sandwich"抽样模型,结合航空影像和TM影像,以山东省细小耕地的调查为例,验证了该抽样模型的可行性,并阐述了其优越性。文献[10]推导了优化的跳批抽样检验模型,在同时考虑生产方与使用方风险的条件下使样本量达到了最小,该模型适合于具有一致质量特性的连续批产品的质量检验。

上述这些抽样检验模型大多是基于传统工业产品建立的,传统抽样检验模型的推导过程中,需根据历史经验提供待检验数据的不合格品率,这些抽样检验理论对于传统工业产品较为实用。而海洋数据因海量、多源、多类和多维等特征,其质量特性具有不确定性现象。传统抽样检验理论不能满足空间数据质量检验的需求。因此,本文提出了基于梯形模糊数的海洋数据质量抽样检验模型,将质量检验问题转化为模糊线性规划模型问题。

2 相关知识

2.1 海洋数据质量检验模型

对待检验的海洋数据质量检验的结果是:该海洋数据批

到稿日期:2014-04-17 返修日期:2014-06-12 本文受国家自然科学基金项目(61272098),上海市自然科学基金(13ZR1455800),国家973项目(2012CB316200)资助。

王振华(1982-),女,博士,讲师,主要研究方向为空间数据的质量控制理论,E-mail:zh-wang@shou.edu.cn;**周雪楠**(1991-),女,硕士生,主要研究方向为海洋空间数据库;**黄冬梅**(1964-),女,硕士,教授,主要研究方向为数据库技术。

为合格数据或该数据为不合格数据。记海洋数据的质量检验模型为 $S(N,n,c)$，其中，N 为数据批的批量，n 为抽样检验所需样本量，d 为样本检验中发现的不合格数，c 为质量判定参数，即接收数。若 $d\leqslant c$，则该批海洋数据视为合格数据；若 $d>c$，则该批海洋数据视为不合格数据。海洋数据质量检验模型的参数定义如下：

常量：N 为海洋数据批量；D 为同一批海洋数据中的不合格品数；d 为被检验样本海洋数据中的不合格品数。

连续变量：$\tilde{p}$ 为待检验海洋数据的模糊不合格品率；p 为实际不合格品率；ε 为某批海洋数据的接收概率残差；α 为在接收质量限下的拒绝概率。

离散变量：n 为进行质量评估所需要的海洋数据样本量；c 为海洋数据批次样本中的评估判定数，即接收数。

用 $L(\tilde{p})$ 表示海洋数据抽样检验模型的接收概率，则其计算公式为：

$$L(\tilde{p})=\sum_{d=0}^{c}\frac{\tilde{\lambda}^{d}}{d!}e^{-\tilde{\lambda}} \tag{1}$$

其中，$\lambda=n\tilde{p}$。

2.2 梯形模糊数

本文将待检验海洋数据批的不确定不合格品率记为 $\tilde{p}$，用梯形模糊数对其进行定义。

定义 1[11] 对于论域 U 上的梯形模糊数 $\tilde{A}$，记为 $\mu_A(x)=(a,b,c,d)$，满足：

1）$\tilde{A}$ 是 U 中的一个模糊子集；

2）对于一个关系函数 $\mu_A(x)=(a,b,c,d)$，其中 x 为变量，a,b,c,d 为常数，且 $a<b<c<d$；

3）$\mu_A(x)$ 在 (a,b) 上单调递增；

4）$\mu_A(x)$ 在 (c,d) 上单调递减；

5）$\mu_A(x)=1$，当 $b\leqslant x\leqslant c$ 时。

隶属函数表示如下：

$$\mu_A(x)=\begin{cases}\dfrac{x-a}{b-a}, & a\leqslant x<b\\ 1, & b\leqslant x\leqslant c\\ \dfrac{d-x}{d-c}, & c<x\leqslant d\\ 0, & \text{other}\end{cases} \tag{2}$$

如图 1 所示，当 $a=b=c=d$ 时，$\tilde{A}$ 转变为确定的实数。

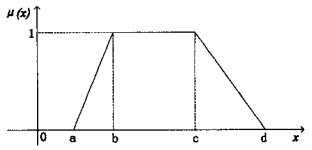

图 1 梯形模糊数示意图

定义 2 对于置信系数 $\alpha\in[0,1]$，当满足如下条件时，$\tilde{A}_\alpha$ 为模糊子集 $\tilde{A}$ 的 α-截集：

$$\tilde{A}_\alpha=[(b-a)\alpha+a,d+(c-d)\alpha] \tag{3}$$

2.3 模糊抽样检验模型的 *OC* 曲线

以不合格品率 $\tilde{p}$ 为横坐标，以质量检验模型的接收概率 $L(\tilde{p})$ 为纵坐标，对于一系列的 $\tilde{p}$ 值，将点 $(\tilde{p},L(\tilde{p}))$ 描绘在坐标平面上，并把这些点用一条曲线连接起来，该曲线称为质量检验模型 $S(N,n,d,c)$ 的抽检特性曲线，简称 *OC* 线[2,12,13]。不同于传统的抽样检验模型的 *OC* 曲线，模糊抽样检验模型的特征曲线为包含了两条 *OC* 曲线的一个条带，简称 *OC-band*，如图 2 所示。

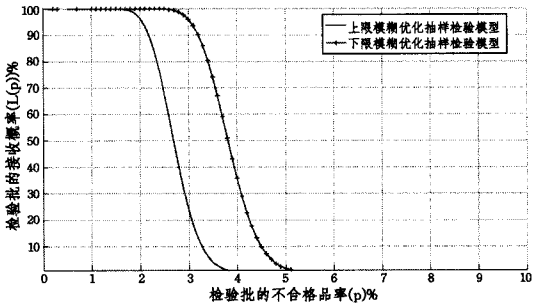

图 2 模糊抽样检验模型的特性曲线

图中，*OC-band* 的上边界线称为上限模糊质量检验模型的 *OC* 曲线；下边界线称为下限模糊质量检验模型的 *OC* 曲线。*OC-band* 的宽度由抽样检验模型的模糊强度确定，例如梯形模糊数中的 a,b,c,d。随着质量参数不确定性的减小，*OC-band* 的宽度也减小；当质量参数的不确定性消失，即模糊参数变为确定参数时，模糊抽样检验模型转化为确定质量参数的抽样检验模型。

3 海洋数据的模糊抽样检验模型

3.1 基于模糊非线性规划算法的海洋数据质量检验模型

在质量检验实施前，给出该数据批的模糊不合格品率 $\tilde{p}_0$。若检验批的不合格品率低于或等于这个值，则该数据批达到质量要求。当检验批的质量水平等于或优于 $\tilde{p}_0$ 时，其判为不合格的概率应不大于 α，即质量检验模型的接收概率不小于 $1-\alpha$。满足该要求的质量检验模型的 *OC-band* 需包含点 $(\tilde{p}_0,1-\alpha)$。通过控制模糊质量检验模型的接收概率上、下限，使其包含点 $(\tilde{p}_0,1-\alpha)$，且模糊质量检验模型中接收数 c 和样本量 n 均为整数，则该海洋数据质量检验的模糊非线性规划模型为：

$$\begin{aligned}&\min_{n}\ \varepsilon^2\\ &\text{s.t. }\{\sum_{d=0}^{c}\frac{n\tilde{p}^{d}}{d!}e^{-(n\tilde{p})}\}-(1-\alpha)=\varepsilon\\ &0\leqslant c\leqslant n-1\end{aligned} \tag{4}$$

式中，N 为批量大小；n 为样本量；$\tilde{p}$ 为模糊不合格品率；ε 为接收概率的残差平方和。

3.2 模糊非线性规划算法

模糊非线性规划算法（Fuzzy Nonliner Programming Algorithm，FNPA）如下：

INPUT：模糊不合格品率 $\tilde{p}_0$，生产方风险 a，批量 N；

OUTPUT：海洋数据质量检验方案 S(N,n,c)。

①初始化：令批量海洋数据样本接收数 c=0，对应的样本量 n=0；

②for (c=0;c<=n;c++){

③ for(n=0;n<N−1;n++){

利用式(4)计算当前状态的残差平方和 ε_c^2；

④ if($\varepsilon_c^2<\varepsilon_{c+1}^2$) break;}}

⑤输出优化方案 S(N,n,c)。

算法分析：

在该算法中，1）求该批海洋数据在给定批量和接收质量

水平下的样本量 n 的时间复杂度为 O(1)；2)求已知当前样本量 n 的接收数 c 的时间复杂度为 $O(N)$，因此该算法的时间复杂度为 $O(N)$；3)在样本量 n 和接收数 c 都未知的情况下时间复杂度为 $O(N^2)$。

4 实验结果与分析

本节以某海域调查数据为例对所提方法的可行性进行验证，通过与传统概率质量检验模型的比较，阐明了该方法的适用条件和优越性。

以某海域调查数据为例，该海域共有 8 个观测站(包括台站、浮标)，各观测站实时提供流速、水温、盐度以及潮汐等空间观测数据，如表 1 所列。因各观测站的设备、技术人员的熟练程度、实际环境等因素的不同，各观测站点所提供数据的质量特性存在较大差异。据历史资料显示(本数据批的前两月数据资料统计)，该海域空间数据的不合格品率大约在 0.02 至0.03 上下波动。在对该海域进行质量检验的过程中，将其不合格品率抽象为梯形模糊数，基于模糊抽样检验模型对其质量精度进行评定。

表 1 观测站某一时刻提供的数据类型

数据类型	流速 (m/s)	水温 (℃)	盐度 (‰)	潮汐 (mm)	气温 (℃)	风速 (m/s)	风向
观测值	4.10	5.98	18.24	374	19.40	10.20	ENE

4.1 模糊抽样检验模型的制定

将不确定的不合格品率抽象为一个模糊数 $\tilde{p}$，则根据梯形模糊数理论，该模糊不合格品率为：

$$\tilde{p}=(0.01,0.02,0.03,0.04) \tag{5}$$

该梯形模糊不合格品率的 α-截集为：

$$\tilde{p}[\alpha]=[0.01+0.01\alpha,0.04-0.01\alpha] \tag{6}$$

基于离散模糊泊松分布，该模糊抽样检验模型的接收概率为：

$$L(\tilde{p})[\alpha]=[L(\tilde{p})^L[\alpha],L(\tilde{p})^U[\alpha]] \tag{7}$$

式中，$L(\tilde{p})$ 为模糊抽样检验模型的接收概率，$L(\tilde{p})^L$ 和 $L(\tilde{p})^U$ 分别为模糊上限和模糊下限的接收概率，由下式计算得出：

$$\begin{aligned} L(\tilde{p})^L[\alpha]&=\min\{\sum_{d=0}^{c}\frac{\tilde{\lambda}^d}{d!}e^{-\tilde{\lambda}}\mid\tilde{\lambda}\in\tilde{\lambda}[\alpha]\}\\ L(\tilde{p})^U[\alpha]&=\max\{\sum_{d=0}^{c}\frac{\tilde{\lambda}^d}{d!}e^{-\tilde{\lambda}}\mid\tilde{\lambda}\in\tilde{\lambda}[\alpha]\}\end{aligned} \tag{8}$$

式中，$\lambda=n\tilde{p}$。

4.2 结果与分析

将观测所得空间数据组批进行质量检验。若数据量过大，则分批次进行质量检验；若空间数据量较小，则合并数据进行质量检验，尽量避免百分比抽样检验中“大批量多宽，小批量过严”的缺陷[14]。在该实验中，分别抽取了待检验空间数据批的 10%、20%、30%作为样本进行质量检验。基于3.1 节推导了不合格品率为梯形模糊数情况下 8 个观测点的模糊抽样检验模型。

表 2－表 4 分别给出了抽样比分别为 10%、20%、30%时的质量抽样检验模型的各参数，其中，N 为批量，n 为样本量，c_1,c_2,c_3,c_4 分别为检验模型的接收数；其中，c_1 为基于梯形模糊不合格品率的上限模糊质量抽样检验模型的接收数；c_4 为基于梯形模糊不合格品率的下限模糊抽样检验模型的接收数；c_2,c_3 为梯形不合格品率转变为两确定数时的概率抽样检验模型的接收数。以观测站 Z_1 为例，图 3－图 5 分别比较了抽样比分别为 10%、20%、30%时，上、下限模糊抽样检验模型与概率优化抽样检验模型的 OC 曲线比较。

表 2 8 个观测点的模糊抽样检验模型($n=N*10\%$)

观测站	批量	样本量	检验模型的接收数			
	N	n	c_1	c_2	c_3	c_4
Z_1	7560	756	12	22	31	40
Z_2	15120	1512	22	40	57	74
Z_3	5040	504	9	16	22	28
Z_4	3780	378	7	12	17	22
Z_5	10080	1008	16	28	40	51
Z_6	7560	756	12	22	31	40
Z_7	6048	605	10	18	25	33
Z_8	4320	432	8	14	19	24

表 3 8 个观测点的模糊抽样检验模型($n=N*20\%$)

观测站	批量	样本量	检验模型的接收数			
	N	n	c_1	c_2	c_3	c_4
Z_1	7560	1512	22	40	57	74
Z_2	15120	3024	40	74	107	139
Z_3	5040	1008	16	28	40	51
Z_4	3780	756	12	22	31	40
Z_5	10080	2016	28	51	74	96
Z_6	7560	1512	22	40	57	74
Z_7	6048	1210	18	33	46	60
Z_8	4320	864	14	24	35	44

表 4 8 个观测点的模糊抽样检验模型($n=N*30\%$)

观测站	批量	样本量	检验模型的接收数			
	N	n	c_1	c_2	c_3	c_4
Z_1	7560	2268	31	57	82	107
Z_2	15120	4536	57	107	145	137
Z_3	5040	1512	22	40	57	74
Z_4	3780	1134	17	31	44	57
Z_5	10080	3024	40	74	107	139
Z_6	7560	2268	31	57	82	107
Z_7	6048	1815	25	46	67	87
Z_8	4320	1296	19	35	49	64

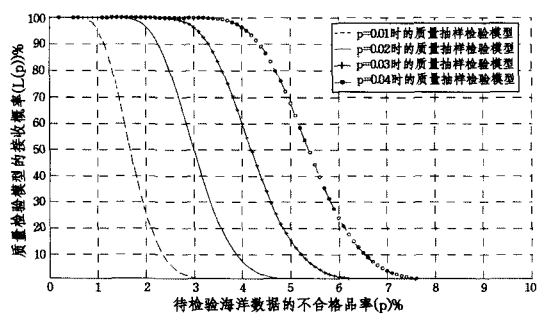

图 3 模糊检验模型和概率检验模型的 OC 曲线比较($n=N*10\%$)

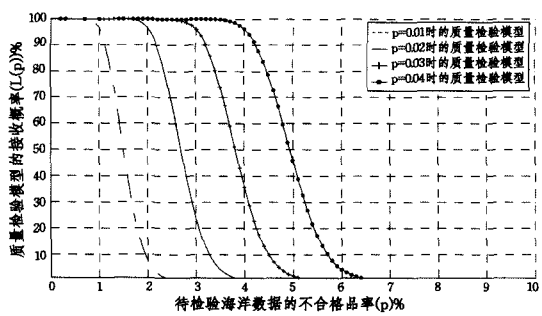

图 4 模糊检验模型和概率检验模型的 OC 曲线比较($n=N*20\%$)

(下转第 190 页)

[15] Elmeleegy H, Elmagarmid A, Lee J. Leveraging query logs for schema mapping generation in U-MAP[C]// Proceedings of the 2011 International Conference on Management of Data. 2011: 121-132

[16] PinkelC. InteractivePayasYouGoRelational-to-OntologyMapping [C]// The Semantic Web-ISWC. 2013:456-464

[17] Aumueller D, Do H H, Massmann S, et al. Schema and ontology matching with COMA++[C]// Proceedings of the 2005 ACM SIGMOD international conference on Management of data. Chicago, IL, USA, 2005:906-908

[18] Peukert E, Eberius J, Rahm E. A self-configuring schema matching system[C]// Proceedings of 28st International Conference on Data Engineering. Washington DC, USA, 2012:306-317

[19] Qian L, Cafarella M J, Jagadish H V. Sample-driven schema mapping[C]// Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. Scottsdale, USA, 2012:73-84

[20] 黄少滨，刘国峰，万庆生，等. 一种基于部分已验证匹配关系的模式匹配模型[J]. 自动化学报，2013，39(10)：1642-1652

[21] 董慧，刘厚嘉. 文献数据库优化设计的探讨[J]. 情报学报，1999，18(1)：43-49

[22] 崔跃生，张勇，曾春，等. 数据库物理结构优化技术[J]. 软件学报，2013，24(4)：761-780

[23] Berzal F, Cubero J C, Cuenca F, et al. Relational decomposition through partial functional dependencies[J]. Data & Knowledge Engineering, 2002, 43(2): 207-234

(上接第 184 页)

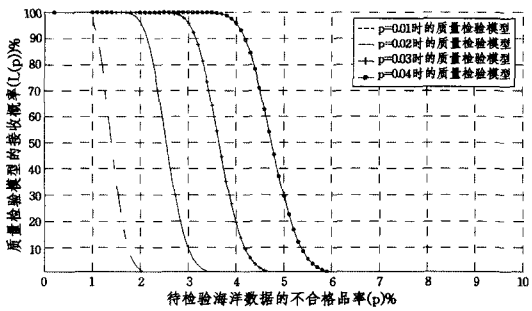

图 5 模糊检验模型和概率检验模型的 OC 曲线比较($n=N*30\%$)

由表 2—表 4 和图 3—图 5 可以看出：1)基于模糊不合格品率可推导出两端点模糊抽样检验模型，即上、下限模糊抽样检验模型。以抽样比为 10%的观测站 1 为例，其上限模糊抽样检验模型为 $S(7560,756,12)$，接收数为 12；下限模糊抽样检验模型为 $S(7560,756,40)$，接收数为 40。即因该海洋数据具有不确定的不合格品率，其质量检验模型的接收数可在 12 至 40 之间选取。2)基于不确定不合格品率的模糊抽样检验模型是具有明确不合格品率质量参数的抽样检验模型的扩充，其可涵盖模糊不合格品率的所有变化情况。即上、下限模糊抽样检验模型的接收数区间涵盖了其不合格品率为确定参数(0.02 或 0.03)时的概率抽样检验模型。3)不同模糊不合格品率的模糊抽样检验模型的辨别率亦不同，即上限模糊抽样检验模型具有最强的辨别力，而下限模糊抽样检验模型的辨别力最弱；用户在不确定不合格品率的情况下，可根据精度要求选择适当的质量抽样检验模型。

结束语 不同于传统工业产品的生产形式，海洋数据的采集方式多式多样，包括实地测量、遥感、摄影测量、数据化、文档报表等，因此海洋数据质量特性具有不确定性。传统的抽样检验方式不能满足海洋数据的质量检验要求。针对海洋数据不确定的质量特性，本文在抽样检验模型的制定中引入了梯形模糊数，扩充了传统概率优化抽样检验模型的制定方法，完善了海洋数据的抽样检验理论体系。

参 考 文 献

[1] 张耀中. 质量抽样检验标准实施指南[M]. 深圳：海天出版社，2004：3-16

[2] 于善奇. 抽样检验与质量控制[M]. 北京：北京大学出版社，1991：15-49

[3] Dodge H F, Roming H G. Single sampling and double sampling inspection tables[J]. The Bell System Technical Journal, 1941, 20(1): 1-61

[4] Dodge, H F. A sampling inspection plan for continuous production[J]. The Annals of Mathematical Statistics, 1943, 14(3): 264-279

[5] Dodge H F, Roming H G. Sampling Inspection Table, Single and Double Sampling[M]. New York: John Wiley & Sons, 1959: 118-220

[6] Jun C H, Balamurali S, Kalyanasundaram M, et al. Evalution and design of two level continuous sampling plans[J]. Tamkang-Journal of Science and Engineering, 2006, 9(4): 409-417

[7] Duarte B P M, Saraiva P M. An optimization-based approach for designing attribute acceptance sampling plans[J]. International Journal of Quality & Reliability Management, 2008, 25(8): 824-841

[8] Eleftherion M, Farmakis N. Continuous sampling plan under quadratically varying acceptance cost[C]// The XIII International Conference "Applied Stochastic Models and Data Analysis". Vilnius, Lithuania, 2009: 289-293

[9] Wang Jing-feng, R Hai-ning, Cao Zhi-dong, et al. Sampling surveying to estimate the mean of a heterogeneous surface: reducing the error variance through zoning[J]. International Journal of Geographical Information Science, 2010, 24(4): 523-543

[10] Aslam M, Balamurali S, Jun C H, et al. Optimal designing of a skip lot sampling plan by two point method[J]. Pakistan Journal of Statistics, 2010, 26(4): 585-592

[11] Ma M, Friedman M, Kandel, et al. A new fuzzy arithmetic[J]. Fuzzy Sets and Systems, 1999, 108: 83-90

[12] Wetherill G B. Sampling Inspection and Quality Control[M]. Chapman and Hall, London, 1977: 233-267

[13] Govindaraju K, Balainurali S. Chain sampling plan for variables inspection[J]. Journal of Applied Statistics, 1998, 25(1): 103-109

[14] 刘大杰，刘春. GIS 数字产品质量抽样检验方案探讨[J]. 武汉测绘科技大学学报，2000，24(4)：348-361