第 38 卷第 5 期
2010 年 5 月

同济大学学报(自然科学版)
JOURNAL OF TONGJI UNIVERSITY(NATURAL SCIENCE)

Vol. 38 No. 5
May 2010

文章编号：0253-374X(2010)05-0749-04

DOI:10.3969/j.issn.0253-374x.2010.05.021

连续大批量空间数据质量抽样检验方案

王振华[1]，童小华[1,2]，梁　丹[1]，谢　欢[1]

(1.同济大学 测量与国土信息工程系，上海 200092；2.同济大学 现代工程测量国家测绘局重点实验室，上海 200092)

摘要：以调整型抽样方案思想为指导，利用连续批的信息量，基于历史资料反泊松推算出接收数 c.同时初步探讨空间数据质量抽样检验中批量 N 与样本容量 n 之间的关系，提出了根据批量 N 和连续提交检查批过程平均上限值 AQL 确定样本容量 n 的计算方法.改进了连续大批量空间数据的抽样方案(N，n，c)的制定方法.最后，通过对改进算法所得抽样方案与百分比抽样方案、查表所得抽样方案进行比较，指出了改进方法的优点及其适用条件.

关键词：质量控制；抽样方案；过程平均上限值(AQL)；抽检特性曲线(OC 曲线)

中图分类号：P 208　　**文献标识码**：A

Sampling Inspection Schemes for Continuous Lot Spatial Data

WANG Zhenhua[1]，*TONG Xiaohua*[1,2]，*LIANG Dan*[1]，*XIE Huan*[1]

(1. Department of Surveying and Geo-informatics, Tongji University, Shanghai 200092, China; 2. Key Laboratory of Modern Engineering Surveying of State Bureau of Surveying and Mapping, Tongji University, Shanghai 200092, China)

Abstract: Based on empirical data, acceptance number(c) is calculated by inverse Poisson. A new method is used to calculate sample size n, which uses lot size N and acceptable quality level of process average (AQL). The scheme of continuous spatial data with large lot size is designed (N, n, c). At last, the spatial data schemes, designed on the basis of the developed algorithm, are compared with those designed on the basis of the percent sampling inspection and table. Comparison results verify the merits and the implementary conditions of the developed algorithm.

Key words: quality control; sampling scheme; acceptable quality level of process average (AQL); operating characteristic curve (OC curve)

抽样检验是按预先确定的抽样方案，从批或过程中随机抽取样本，逐个检验样本，并对批或过程质量做出是否接收的判定，是介于不检验与百分比检验之间的一种检验方法[1].科学的抽样检验方法已有 70 多年的历史，通过生产实践，国际上陆续制定了一些抽样标准如美国军用标准 ISO 2859、基于地理信息的 ISO 19113 和 ISO 19114 等[2-4].同时我国根据检验产品的不同也制定了相应的标准，如基于不合格品率的抽样表、孤立批计数抽样表、数字测绘产品检验验收规定和计数抽样表等[5-8].这些抽样标准的制定不仅使抽样理论形成了完善的体系，同时也规范了抽样检验程序.目前，很多学者将传统的抽样理论体系应用于空间数据产品的检验和调查，例如 Gillies 将 MIL-STD-1916 和 ISO 19114 相结合，提出适合于空间数据产品的检验流程[9]；Brus 和 Banjevi 等基于空间关联性优化了样本点的选择[10-11]；王劲峰等建立了“sandwich”模型，对耕地中细小的土地利用变化进行了抽样检测[12]；此外还有很多学者将空间抽样理论应用于土地调查、环境污染、房产测量成果检查验收等方面[13-15].

与传统独立的工业产品相比，空间数据产品具有自相关性及变异性等特点，且表达形式多样.因此传统的抽样检验理论直接用于空间数据的质量控制存在一些不足，如缺少规范、统一的空间数据质量模型；空间数据批量 N 与样本容量 n 之间缺少科学依据；连续空间数据的信息量没有得到充分利用等，同时空间属性数据的定性化表述也不利于抽样检验的实施.

本文从抽样检验方案的接收概率出发，以计数调整型抽样方案思想为指导，对连续大批量空间数据抽样方案的制定提出了改进算法，最后通过与百分比抽样方案和查表抽样方案的比较，指出了改进后抽样方案的优点.

收稿日期：2009-02-24

基金项目：国家自然科学基金资助项目(40771174)；高等学校博士学科点专项科研基金资助项目(20070247046)；上海市青年科技启明星计划(跟踪)资助项目(08QH14022)；上海市曙光计划资助项目(07SG24)

作者简介：王振华(1982—)，女，博士生，主要研究方向为空间数据质量控制与抽样检验. E-mail:wangzhenhua0531@126.com

童小华(1971—)，男，教授，博士生导师，工学博士，主要研究方向为遥感与空间数据处理. E-mail:xhtong@tongji.edu.cn

1 空间数据产品质量抽样检验的接收概率

对一批产品进行抽样检验时,其检验结果可能被接收,也可能被拒绝.设产品的批量为 N,不合格品率为 p,接收概率与不合格品率 p 有关,记为 $L(p)$.抽样方案记为$(N;n,c)$.其中,n 表示抽样的样本容量;c 表示合格判定数,即接收数.当检验后查出不合格品数为 d,则判定规则是:当 $d=c$ 时接收此批产品;当 $d>c$ 时则拒收此批产品,其中 d 为一随机变量.

对于大批量的空间数据质量产品,若其批量为 N,不合格品数 $D=Np$,从 N 件中任取 n 件,当$n/N<0.1$,$p<0.1$ 时,可用泊松分布计算接收概率,其中 $\lambda=np$.则恰好有 d 件不合格品的接收概率为[16]

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

2 空间数据产品质量的抽样方案

抽样方案是规定样本容量 n 和接收准则的具体方案[16].空间数据产品通常以数字地图或空间数 P 据库的形式存在,其生产通常采用统一的数据产品规范,每一个负责数据生产的单位生产过程连续且稳定,但是数据质量有好有坏[17].当对空间数据产品进行抽样检验时,首先需对产品进行质量元素界定,基于空间数据质量评价方法将类型复杂多样的检查项定量化为统一量纲,最终将产品以连续批或单批的形式提交.

根据空间数据产品的特点,应用计数调整型抽样方案思想,在连续提交检查批的过程平均上限值(AQL)确定的情况下,根据统计学理论确定批量 N 与样本容量 n 之间的科学关系;充分利用连续批的信息量,利用历史数据反泊松推算出接收数 c,制定连续大批量空间数据的抽样方案(N,n,c).

2.1 接收数 c 的确定

根据抽样检验的判定规则,从批中抽取 n 件产品构成样本容量,然后逐个检验样本,发现 d 件不合格品,若 $d<c$,则该批产品判断为合格;若 $d>c$,则该批产品判断为不合格.c 为抽样方案中的接收数.由式(1)可知,当批量 N 非常大,且 $n/N<0.1$,$p<0.1$ 时,检验产品的接收概率由泊松分布计算得出.反之,因为连续批生产通常采用统一的数据产品规范,每一个负责数据生产的单位生产过程连续且稳定,连续批相邻两批产品其质量水平具有强相关性.将上一批不合格品率 p'作为历史数据(可看作本批产品不合格率的估值),在样本容量 n 以及质量水平 AQL 确定的条件下,给定预期的接收概率,反泊松推算抽样方案中接收数 c.

已知 AQL 以及样本容量n,将历史数据 p'作为本批产品不合格率的估值,使式(2)得到最优解的 c 就是抽样方案的接收数,如

$$\min\left(\sum_{d=0}^{c}\frac{\lambda}{d!}^{d}\mathrm{e}^{-\lambda}-\mathrm{Pr}\right)^2<\varepsilon \tag{2}$$

式中:Pr 为接收概率(可取一段区间);ε 为任意小.由式(2)可以看出,因抽样方案中接收数为整数,故可推算出最优的接收数 c;不合格率 p'与检验批抽样方案中的接收数 c 成负相关,即当连续批质量变坏时,不合格率 p'增大,则抽样方案中接收数 c 减小,则抽样方案自动调整为加严检验.

2.2 抽样检验特性曲线(OC 曲线)

抽样检验特性曲线(OC 曲线)是批接收概率 $L(p)$随批的质量不合格率 p 而变化的函数曲线,是建立和选择抽样方案的依据之一.根据技术条件或供货方合同,假定质量标准为 p_0,则抽样方案的辨别率是指,对 $p<p_0$ 的高质量产品以低概率拒收(保护生产方);同时对 $p>p_0$ 的低质量产品以高概率拒收(保护使用方)的综合能力.辨别率指数以 OR 表示,即接受概率为 10 与 95%时对应产品质量的比值,OR 越小其所对应抽样方案辨别率越强[16-18].

3 实例分析

以某市空间地质数据库为例,检查项包括坐标系或投影的正确性、空间实体位置的准确性、校正控制点分布的合理性、图层的完整性、拓扑关系的正确性等.记录检查中缺陷个数,以缺陷率作为其不合格品率.根据数据库的生成条件及使用方的精度要求,规定过程平均上限值 AQL 为 3,据历史资料表明其不合格品率 p 在 1%左右浮动,且 $\alpha=0.05$ 的置信度水平下要求抽样最大相对误差不超过 $d=0.2$.

3.1 抽样方案制定

根据检验要求,本文制定了三种不同的抽样方案:传统百分比抽样方案、查 GB 2828 表抽样方案以及改进算法抽样方案.传统的抽样方案中,大多采用百分比的方法确定样本容量 n,其缺点是“大批量过严,小批量过宽”[19].随着抽样理论的发展,国内外制定了各种抽样标准,使得样本容量 n 可方便地通过查找这些抽样表获取[2-8].但标准的制定大多采

用优先数确定样本容量 n 与批量 N 之间的关系[8]，缺少严格的理论根据，同时批量按范围进行了分割，缺乏灵活性.

在改进算法计算抽样方案中，首先确定样本容量 n，抽样中要控制抽样误差，需规定可允许最大误差. 对于不合格品率 p 的估计量 $\widetilde{p}$，可提出如下精度要求，给定置信水平 $1-\alpha$，允许 p 的最大相对误差为 d [16]，即

$$P_r\{p - \widetilde{p}/p < d\} = 1 - \alpha \tag{3}$$

式中：P_r 同式(2)为接收概率. 当批量足够大时，可认为 $\widetilde{p}$ 服从正态分布. 连续大批量空间数据，其实际批不合格率 p 不可获得，若以 A_{qL}（每百单位产品中不合格品数的上限值，AQL）代替之，根据式(3)可估算样本容量 n，即

$$\widetilde{n} = \frac{\mu_{1-\frac{\alpha}{2}}(1 - A_{ql} \cdot 100\%)}{d^2 A_{ql} \cdot 100\%} \tag{4}$$

$$n = \frac{\widetilde{n}}{1 + \dfrac{\widetilde{n} - 1}{N}} \tag{5}$$

由式(4)，(5)可以看出，在 A_{ql} 固定时，样本容量 n 是最大相对偏差 d 的递减函数，其取值可根据产品的复杂程度和精度要求而定. 确定样本容量后，可根据式(2)计算抽样方案中的接收数 c，以此确定抽样方案.

3.2　结果分析与比较

通过传统百分比抽样，查 GB 2828 表以及改进算法三种方法制定不同质量批下的抽样方案表（n，c），表 1 为各抽样方案的批量（N），样本容量（n），接收数（c），抽样比（f）和抽样方案的辨别率（OR）. 选取三种方法中批量为 500，5 000 和 30 000 所对应的抽样方案，应用 Matlab 7.1 画出各抽样方案对应的 OC 曲线，如图 1.

表 1　不同批量抽样方案表

Tab.1　Sampling schemes for different lots

N/个	百分比抽样方案				查表抽样方案(GB2828)				改进算法抽样方案			
	n/个	c/个	f/%	OR	n/个	c/个	f/%	OR	n/个	c/个	f/%	OR
200	20	0	10	46.00	32	2	16.00	6.99	188	3	94.00	5.00
500	50	1	10	10.71	50	4	10.00	4.00	431	6	86.20	3.20
700	70	1	10	11.00	80	6	11.43	3.25	571	8	81.57	2.81
1 000	100	1	10	10.86	80	6	8.00	3.25	757	10	75.70	2.63
2 000	200	1	10	11.61	125	9	6.25	2.69	1 217	16	60.85	2.00
5 000	500	1	10	13.64	200	12	4.00	2.36	1 916	24	38.32	1.89
7 000	700	1	10	11.00	200	12	2.86	2.36	2 151	27	30.73	1.80
9 000	900	1	10	10.63	200	12	2.22	2.36	2 309	28	25.66	1.76
11 000	1 100	1	10	8.75	315	18	2.86	1.98	2 422	30	22.02	1.76
30 000	3 000	2	10	6.80	315	18	1.05	1.98	2 814	34	9.38	1.67

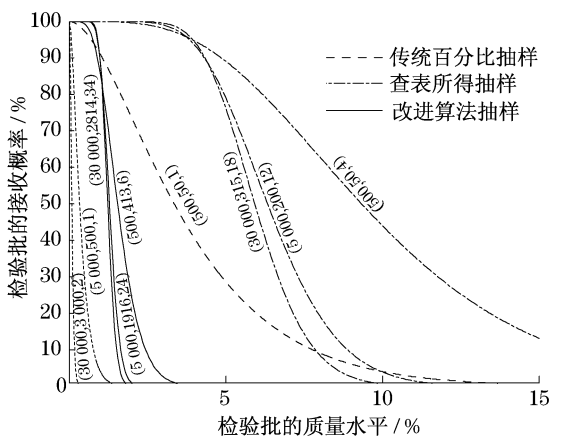

图 1　各抽样方案 OC 曲线表

Fig.1　OC curves of sampling schemes

由表 1 和图 1 可以看出，百分比抽样方案存在“大批量过严，小批量过宽”的缺点. 查表或改进算法所得的抽样方案随批量 N 的增大，接收数 c 越来越大，而抽样比越来越小，克服了传统百分比抽样中的缺点. 通过 OR 的比较，可看出当批量较少时，改进算法所得抽样方案接近或优于查表所得抽样方案，但其抽样比较大，接近于全检，故对于小批量数据改进算法对抽样方案的优化效果不明显. 由于对批量范围的割分，查表所得抽样方案出现“不同的批量对应同一抽样方案”，而改进算法根据批量的不同，制定出更高效的抽样方案，同时改进算法是在经验数据，以及给定接收条件计算抽样方案，所以当批量达到一定的值（如表 1 中 $N \geqslant 5\ 000$），其不管批量如何增大抽样方案的辨别率相近，故消除因批量大小造成生产方或使用方利益问题.

4　结语

抽样检验多是基于传统的独立样本或生产过程，

而空间数据类型多种多样,将传统的抽样检验理论直接应用于空间数据产品,无论在样本容量的确定还是在接收准则的制定方面都存在明显的不足.本文在计数调整型抽样方案制定思想的指导下,提出了根据接收概率函数反算出接收数 c;并且根据批量 N 和过程平均质量上限 AQL 确定样本容量 n 的方法,从而改进了抽样方案的制定.在实例分析中,分析比较了传统百分比抽样方案、查表 GB 2828 抽样方案与改进算法所得抽样方案.结果表明,改进算法所得抽样方案不仅克服了传统百分比抽样方案"大批量过严,小批量过宽"的缺点,而且具有较好的理论依据及较强的辨别率,既保证了使用方的权益,又不损害生产方的利益,且不同的批量对应不同的抽样方案,比查表所得抽样方案具有更强的灵活性.

考虑到空间数据产品涉及的内容和质量特性较为复杂,同时具有计数质量特性和计量质量特性的特点.抽样检验模型还有待更为深入地研究,从而满足不同的抽样水平,适合复杂程度不同的产品,在抽样费用最小的情况下保证抽样的可靠性.

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.
ZHANG Yaozhong. Standard guidelines of quality sampling inspection[M]. Shenzhen: Seasky Publishing House, 2004.

[2] International Organization for Standardization. ISO 2859—1. Sampling procedures for inspection by attributes. Part 1: Sampling schemes indexed by acceptance quality limit _AQL_ for lot-by-lot inspection[S]. Geneva: ISO, 1995.

[3] International Organization for Standardization. ISO 19113. Geographic information—quality principles [S]. Geneva: ISO, 2002.

[4] International Organization for Standardization ISO 19114. Geographic information—quality evaluation procedures [S]. Geneva: ISO, 2003.

[5] 国家质量技术监督局.GB/T 13262—91 不合格品率的计数标准型一次抽样检查程序及抽样表[S].北京:国家质量技术监督局,1992.
Institute of Standardization for Surveying and Mapping. GB/T 13262—91 Single sampling procedure and tables for inspection having desired operating characteristics by attributes for percent nonconforming[S]. Beijing: State Bureau of Quality and Technical Supervision, 1992.

[6] 国家质量技术监督局.GB/T 15239—94 孤立批计数抽样检验程序及抽样表[S].北京:国家质量计数监督局,1995.
Institute ofStandardization for Surveying and Mapping. GB/T 15239—94 Sampling procedures and tables for isolated lot inspection by attributes[S]. Beijing: State Bureau of Quality and Technical Supervision, 1995.

[7] 国家质量技术监督局.GB/T 18316—2001 数字测绘产品检查验收规定和质量评定[S].北京:国家质量技术质监局,2001.
Institute ofStandardization for Surveying and Mapping. GB/T 18316—2001 Specifications for inspection, acceptance and quality assessment of digital surveying and mapping products [S]. Beijing: State Bureau of Quality and Technical Supervision, 2001.

[8] 肖惠,张玉柱,马毅林,等.GB/T 2828.1—2003 计数抽样检验程序第1部分:按接收质量限(AQL)检索的逐批检验抽样计划"理解与实践[M].北京:中国标准出版社,2003.
XIAO Hui, ZHANG Yuzhu, MA Yilin, et al. GB/T 2828.1—2003 Sampling procedures for inspection by attributes: part 1 sampling schemes indexed by acceptance quality limit (AQL) for lot-by-lot inspection. general administration of quality supervision [M]. Beijing: Standards Press of China, 2003.

[9] Gillies F. Geospatial statistical quality management: integration of MIL-STD-1916 and ISO 19114: 2003 [C] // ESRI User Conference Proceedings, San Diego: [s.n.], 2007: 1-13.

[10] Brus D J, de Gruijter J J. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil(with discussion)[J]. Geoderma, 1997, 80: 1.

[11] Banjevic M. Optimal network designs in spatial statistics[D]. Stanford: Stanford University. Department of Statistics, 2004.

[12] WANG Jinfeng, ZHUANG Dafang, LI Lianfa. Spatial sampling design for monitoring the area of cultivated land [J]. International Journal of Remote Sensing, 2002, 13(2): 263.

[13] Arbia G, Lafratta G. Anisotropic spatial sampling designs for urban pollution[J]. Journal of the Royal Statistical Society, Series C-Applied Statistics, 2002, 51: 223.

[14] Lark R M. Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood[J]. Geoderma, 2002, 105(1-2): 49.

[15] 余晓红.建设管理地理信息系统数据的质量控制——以上海市浦东新区为例[D].北京:中国科学院,2002.
YU Xiaohong. Theories and methods of data quality control in construction and management GIS——taking Shanghai Pudong GIS as Practice [D]. Beijing: Chinese Academy of Sciences, 2002.

[16] 于善奇.抽样检验与质量控制[M].北京:北京大学出版社,1991.
YU Shanqi. Sampling inspection and quality control [M]. Beijing: Peking University Press, 1991.

[17] 国土资源部.国土资源数据质量检查与抽样检验[R].北京:国土资源部,2008.
Ministry of Land and Resources of P R C. Quality inspection and sampling inspection of land resources data[R]. Beijing: Ministry of Land and Resources of P R C, 2008.

[18] 张玉柱,曹世民,胡自伟,等.调整型抽样检验系统理论与应用[M].北京:国际工业出版社,2005.
ZHANG Yuzhu, CAO Shimin, HU Ziwei, et al. The theory and application of adjusting attribute sampling system[M]. Beijing: National Defense Industry Press, 2005.

[19] 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,2000,4(25):348.
LIU Dajie, LIU Chun. Study on sampling inspection schemes to digital products in GIS [J]. Journal of Wuhan Technical University of Surveying and Mapping, 2000, 4(25): 348.