# 地理数据空间抽样模型*

李连发 王劲峰

中国科学院地理科学与自然资源研究所资源与环境信息系统国家重点实验室，北京 100101

**摘要** 提出了空间抽样的一般模型及其构造应用模型的框架；并采用组件对象模型方法编制了基于一般模型有较强通用性的空间抽样集成(SSI)软件包. SSI在自然灾害监测及可耕地面积调查的实际应用中取得了很好的结果，显示出一般模型及其应用框架与COM集成软件包思路的优点.

**关键词** 空间抽样 地理信息 一般模型 软件集成

抽样是将总体集(连续)按某种规律划分为样本集(离散)且不损失总体主要信息的方法；空间抽样则是针对在地理空间上分布且相互间有关联性的研究对象而言，它是具有空间关联性的抽样. 合理的抽样是一种经济、快速、及时、质量好及准确性高的调查方式[1,2]. 高效率的抽样也可看作是一个方案的优化选择问题.

在生态环境监测、自然灾害预报、国土监测、农情速报及社会经济调查等若干国家重大项目的实际应用中，抽样充分显示了巨大的经济效益. 人们已达成这样的共识：空间数据的采集是地学分析的第一步，而空间数据的不同采样策略对最终结果有很大影响. 因此，空间抽样模型的建立将为诸如大范围分区调查、海洋检测、多项目综合调查和动态检测提供科学依据，实现快速、低费和准确地监测和调查，为国家重大项目实施提供支持[3].

简单回顾空间抽样模型形成及发展，它大致经历了经典抽样、连续及离散关联的应用模型阶段，但其基本模型的建立及发展尚处于初步阶段，目前对这一重要问题的综合研究尚不完善，缺乏相应的集成软件包. 本文在实际工作基础上对此进行了探索.

## 1 一般模型及集成软件包

抽样是一个普遍但受多因素影响的复杂问题，空间抽样问题的本质简单描述为：在现有经济条件下，采用怎样的抽样方案使得从有限个样本值得到的估计值更接近总体真值，或者说怎样使精度与代价之间达到最优平衡.

在对典型的抽样及空间抽样应用模型进行研究及总结后，结合抽样本身的特性，可将空间抽样问题分解成3个方面：(1) 样本选取方式(包括简单随机、分层、系统、整群及多阶段等)；(2)空间关联性，即如何确定样本点之间存在的空间关联特性(无空间关联、连续关联或离散关联)；(3)精确度，可用方差(误差)的倒数来衡量，是决定抽样方法好坏的标准，同时也是作样点数-精度图用于抽样决策的关键. 问题的解决一般都从这3方面入手，分别找到满意的方法，从而使整个问题得以解决.

针对上述3点，我们提出了解决典型空间抽样问题的一般模型，它提供了解决问题的基本思路；并采用组件对象模型，编制出空间抽样集成软件包SSI进行实现.

### 1.1 一般模型

一般模型的基本框架如下：

(1) 模型输入

① 确定抽样单元数

先验值：总体单元数(只限于有限总体)或抽样比，元素方差 $S^2$ 的估计等；

期望值：误差限，统计推断置信度 $1-p$；

方法：方差上限、绝对误差限或相对误差限确定(由误差限类型确定)[4].

2001-05-14 收稿，2001-07-18 收修改稿

* 中国科学院项目(KZ951-A1-302、KZ951-A1-203、KJ951-B1-703)，国家自然科学基金(批准号：49871064、69896250)和地理科学与资源研究所课题(CX10G-D00-02)共同资助

E-mail：spatialinfo@sina.com.cn

② 样本数(可从①得到)、样本值(空间抽样包括位置及属性值)及样本出现的概率(等概率时为等概率抽样).

(2) 模型输出

① 估计值：估计总体均值，估计均值的方差估计值，标准差；

② 统计推断：置信区间(估计均值以 $1-p$ 的概率位于估计均值的区间内)；

③ 抽样设计依据：样本数-方差图(在无偏情况下，相当调查费用-误差图).

(3) 基本计算式

基本计算式分为 Kriging 方法参见文献[5～8]；空间关联的统计学方法参见文献[1, 4, 9, 10].

在无偏情况下，方差是衡量精度的关键，也是抽样设计效果的中心问题.

对样本数-方差图，分别取不同的样本数，并取样计算均值及方差，绘得函数图；最优化抽样方案，可从该图中求出(一定费用相对的最小方差或一定方差对应的最小抽样调查费).

(4) 模型组成及应用框架

一般模型由选样方式、空间相关性及精确度3部分组成. 各部分组成及其适用情况参见表1.

**表1 一般模型的基本组成及适用情况**

| 组 成 | 方 法 | 特 点 | 适用情况 |
|---|---|---|---|
| 选取样本方式 | 简单随机 | 随机等概率抽取样本. 最基本选样方法. | 空间样本点均匀分布、变化平稳. |
| | 分层 | 将总体划分成若干相互独立的子总体(层)，各层加权估算总体的值. 有比例分层、最优化分层及按指标分层，视应用目标而定. | 分层时层内变差小而层间变差大. 较大区域抽样的基础，如森林调查[11]、作物估产[12]. |
| | 系统 | 将总体单元按 某种顺序排列，在规定的范围内随机抽区起始单元，然后按一套规则抽取其他样本. | 抽样框实施简便易行；精度与总体的排列顺序有关(线形时提高精度；周期变化时与初始点及间距有关)；用在监测网络或格网取样中. |
| | 整群 | 将总体按一定的指标划分成若干子总体的方法. 各子总体分别抽样. | 总体构成复杂，组成总体元素组之间差别较大. 便于大范围的抽样调查. |
| | 多阶段分层[2,13] | 将总体按一定的指标划分成若干子总体，子总体划分成更小的子总体；直至能直接实施元素调查. | 在大范围调查，元素组成复杂且分层明显的样本. 如全国土地抽样调查，按省、区、县、镇及乡的五级分层. |
| 空间关联性 | 连续地物关联[9] | 空间关联函数(指数与协方差模型[9]). | 降雨等空间上连续地物关联函数计算. |
| | 离散地物关联[8,10] | 离散地物空间关联函数的计算，扩大了空间抽样方法的应用对象. | 耕地等离散且具有较强的空间关联性的地物的抽样调查. |
| | 关联函数矩阵模型[14] | 与3个模型 SAR，CAR 及 MA 相对的空间关联矩阵；用最大似然法 ML 估算. | 遥感图像处理中(如进行环境遥感监测)等. |
| 精确度(方差) | 数理统计方法 | 经典统计与空间关联性相结合的统计学的方法. | 具有空间关联性的大多数情况(经典统计学方法). |
| | Kriging 抽样[5～8] | 通过模拟连续区域实现对离散区域的优化抽样决择. | 环境监测、遥感精度评估、土壤污染测定、寻找矿产资源等. |

解决具体问题时，应根据选样方式、空间关联性及精确度对抽样方案进行评估. 结合对象及目标的特点选择适当的方法.

以中国洪水灾害监测网络设计为例，由于全国灾害监测站点一般要求分区均匀布设，因此选择了分层抽样框(相当于分区)，而每层内采用系统或随机抽取样点的方法；由于降水已有连续关联函数模型[9]，且关联性不可忽略，故相关性采用 Bessel 或指数模型；精确度衡量采用与空间关联结合的统计学方法. 为设计采样监测网络，需要进行点探测，作样点数-精度图，便于选择样本设计方案. 具体算例参见2.1节.

### 1.2 集成软件包 SSI

我们用组件对象模型 COM 来编制空间抽样集成软件包 SSI. COM 是从 ActiveX 的基础上逐步发展形成的，“组件”指实现某项功能的独立模块，它的接口是多重继承，故各“组件”可以实现与原有应用兼容下的不断升级[15]，各组件可重新组合. COM 在软件集成方面功能强大；目前已有许多功能强大的统计及地理信息系统(GIS)组件模块供选用，在建立应用空间抽样模型时即可将空间抽样模型相应组件，建立专门的空间抽样系统，也可通过接口嵌入到其他系统之中去作为一子功能. 这种集成方式灵活多变，为一般模型扩展了广阔的应用空间. 而 SSI 的实现，由以下两方面构成.

(1) 模型集成 SSI 是按照一般模型的组成及其结构来进行集成的，SSI 中主要的组件相当于一般模型的基本组成部分，而 SSI 解决实际抽样问题

结构相当于一般模型的应用框架. SSI 既可看成单一的空间抽样系统，它分离的各组件又可看作实现应用空间抽样模型的供选部件集. SSI 组件集成为分 4 层的树状结构(图 1). 在构造应用模型时，只根据需要挑选相关的组件“部件”，而不需要将整个组件都包括，实现了“即插即用”的集成思路.

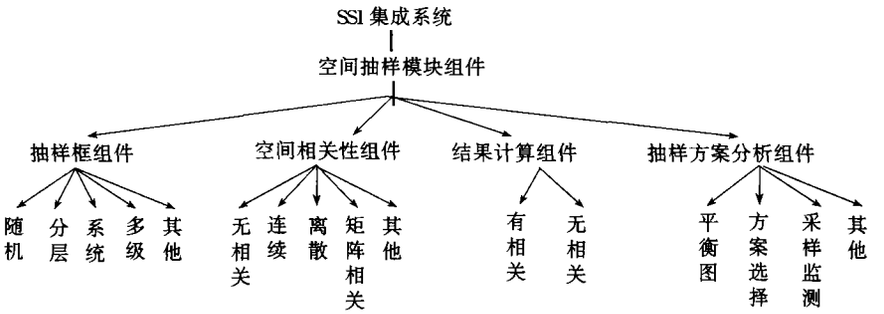


**图 1 采用组件集成 SSI 的树状结构图**

(2) 系统 可将各组件结合起来，并充分应用 GIS, MatLab 及 SAS 等功能组件，形成功能强大的空间抽样系统. 同时，结合应用目标及实际情况，挑选适当的组件组成应用抽样模型；也可嵌入到其他系统中去.

## 2 应用模型实例

### 2.1 中国干旱洪水地震灾害监测

将一般模型及其集成系统 SSI，应用于中国干旱洪水地震灾害数据，进行监测网络设计[16]，构成了应用模型 DMM.

在选样方式、空间相关性及精度衡量方面选用的情况及原因参见 1.1 节中一般模型的灾害监测实例分析，据此选用的组件构成参见图 2. 采样方案的分析是为应用目标－采样监测服务，首先要进行一系列的空间布点探测分析，求得相应的精度－样点数图(图 3)，为实现平衡优化方案的选择提供依据；由此选用的组件有：平衡图分析及方案选择组件等.

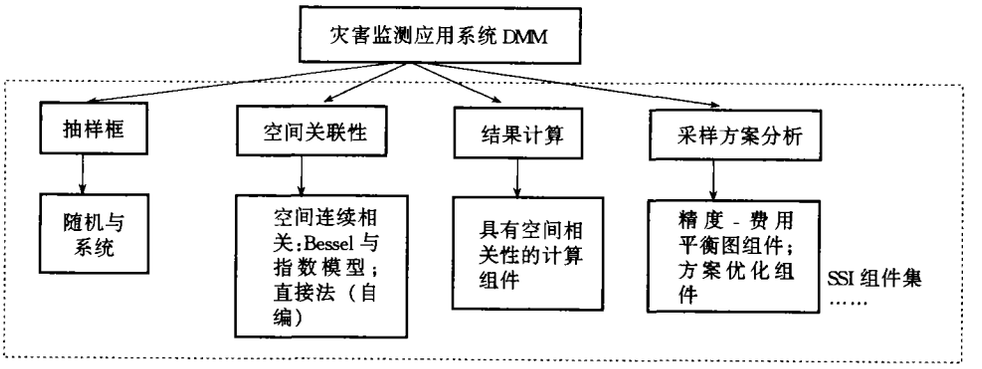


**图 2 DMM 组件构成**

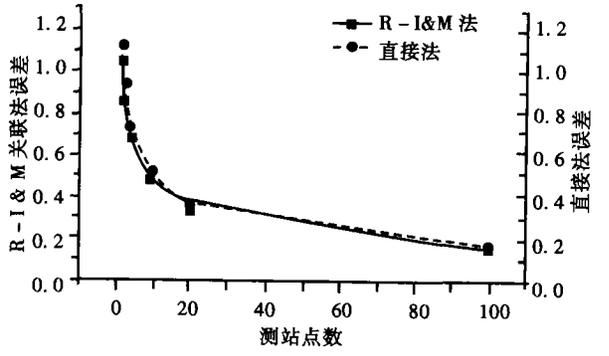


**图 3 由 DMM 所形成的站点数-变差之间函数关系**

图 3 为灾害监测中测值的绝对变差与站点数之间的递减关系图. 它对于优化抽样方案选择的意义在于，当布设的样点(站点)数增加时(经费也相应增加)，变差变小，精度增加. 当样点数增到一定数目时，精度增加幅度变得越来越小，此时追加经费增加监测站点对增加精度已没有多大意义，要根据站点数-变差图选择站点数与要求精度之间的最佳组合.

### 2.2 非耕地面积抽样调查

用一般模型及其 SSI 我们构成了非耕地面积抽

样调查应用模型(NFSSM)[10]. 基本思路同2.1节, 应用结构类似图2. 该模型的抽样框采用了分层与随机组件, 空间相关性用离散地物相关性计算组件(对耕地调查); 对于精度评估, 同样用到了平衡图分析及方案选择组件. 用专题制图仪(TM)影象对山东省非耕地面积动态抽样调查计算结果制成样点数-非耕地比例图(图4). 图中, 非耕地比例是相对于总面积而言. 此图说明当样点增加到一定数目, 计算出的非耕地比例值趋于稳定, 而变差变小且变化趋于缓和(精度增加且也趋于缓和). 它对于抽样方案选择的意义同2.1节.

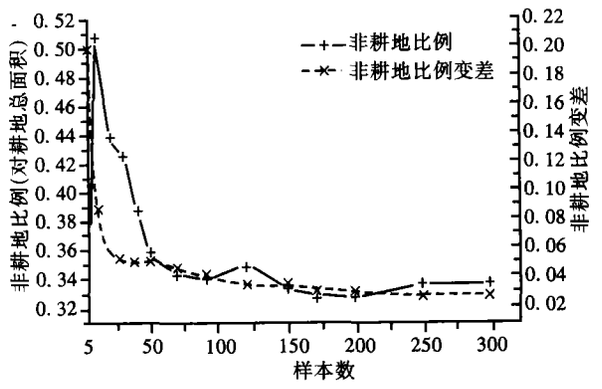


**图4 样点数-非耕地比例图(分区法)**

## 3 讨论与结论

尽管一般模型及其集成软件包SSI提供了针对典型空间抽样问题的一般解决方法, 但仍有若干难题. 比较典型的有动态监测及预报模型中的时间插值与抽样问题[10], 抽样前先验数据的获取及样本的代表性[17]的问题, 目前还没有有效的解决方法.

空间抽样应用性很强, 具有很强的交叉性及多学科性. 概率论、数理统计、经典抽样理论及Kriging方法等奠定了其理论及方法基础, 时间序列分析为其提供了动态监测中的时间抽样工具[10], 误差分析及插值理论则提供了提高精度的工具. 同时, 类似的采样相关问题及其好的解决方法在图像处理、图形压缩、模式识别等计算机学科中很普遍(如图像压缩中将图像进行多分辨率分解的小波变换法). 这些领域中的信息抽样方法能否用到地理信息的抽样策略中, 是一个区别与结合的问题.

总之, 空间抽样问题的解决要求注重它的交叉性及多学科性, 在应用中不断总结与发展相关理论及方法; 注重计算机等学科中信息抽取方法的优点及适宜性. 同时应将好的新方法编制成组件, 加入集成系统中, 对其进行扩充, 以形成通用性强的空间抽样软件包.

## 参 考 文 献

1 Cochran W G. Sampling Techniques. 3d ed. New York: John Wiley & Sons, 1977. 1~355
2 Kish L. Survey Sampling. New York: John Wiley & Sons, 1985. 1~327, 644~691
3 王劲峰, 等. 地理信息空间分析的理论体系探讨. 地理学报, 2000, 55(1): 92
4 冯士雍, 等. 抽样调查-理论、方法与实践. 上海: 上海科学技术出版社, 1996. 1~200
5 Atkinson P M, et al. Exploring the relation between spatial structure and wavelength: Implications for sampling reflectance in the field. International Journal of Remote Sensing, 1999, 20: 2663
6 Atkinson P M. Geographical information science: Geostatistics and uncertainty. Progress in Physical Geography, 1999, 23: 134
7 Haining R. Spatial Data Analysis in The Social And Environmental Sciences. Cambridge: Cambridge University Press, 1991. 171~195
8 Journel A, et al. Mining Geostatistics. London: Academic Press Inc, 1978 . LTM, 1~300
9 Ignacio R I, et al. The design of rainfall networks in time and space. Water Resources Research, 1974, 10: 713
10 Wang J, et al. Spatial sampling design for monitoring the area of cultivated land. International Journal of Remote Sensing, 2002 23(2): 263
11 赵宪文. 林业遥感定量估测. 北京: 中国林业出版社, 1997. 1~300
12 吴炳方. 全国农情检测与估产运行化遥感方法. 地理学报, 2000, 55(1): 25
13 柯惠新, 等. 中国人民银行城镇户抽样调查方案的设计. 数理统计与管理, 1999, 6 (18): 45
14 Haining R. Estimating spatial means with an application to remote sensing data. Communication Statistics -Theory Meth, 1988, 17 (2): 537
15 Dale R. Inside Com. Washington: Microsoft Press, 1997. 1~400
16 王劲峰, 等. 中国干旱洪水地震灾害监测空间采样设计. 自然科学进展, 1999, 9(4): 336
17 Cressie N. Statistics for Spatial Data. New York: Wiley & Sons, 1991. 21~325