【理论与方法】

地理坐标信息参与下的空间平衡抽样设计

郝一炜　　金勇进

【摘　　要】地理信息系统的广泛使用为抽样调查中引入了总体单元的空间信息,随之产生的空间相关性破坏了总体单元之间的独立性假设,传统抽样方法在空间总体中的应用也面临着样本代表性下降的困扰。针对这一问题,提出了利用地理坐标信息选取空间平衡样本以提高样本代表性,并利用样本中包含的地理坐标信息改进方差估计量以提高估计效率。模拟研究和实证分析结果表明,基于空间平衡样本的统计推断在空间相关性较强时能够显著提高估计效率。

【关 键 词】地理坐标;空间信息;空间平衡样本;空间相关性

【作者简介】郝一炜,首都医科大学附属北京地坛医院(北京　100015),北京市卫生健康委员会医疗管理数据质量控制与改进中心(北京　100035);金勇进(通讯作者),中国人民大学应用统计科学研究中心,中国人民大学统计学院,E-mail:jinyongj_519@aliyun.com(北京　100872)。

【原文出处】《数理统计与管理》(京),2020.6.978~989

【基金项目】国家社科基金项目(15BTJ014);中国人民大学双一流建设项目。

0　引言

随着地理信息技术的进步,越来越多的地理学工具被应用在抽样调查的实际工作中。地理信息系统(GIS)、全球定位系统(GPS)和遥感技术(RS)能够获取地表指定单点处的经度、纬度和海拔,各类网络地图的发展完善能够方便的查询企业、学校、实体商户、住宅区等任何具有实际空间坐落的单元的地理信息,为抽样调查中提供了新的信息来源。我们将诸如行政区划、经纬度和海拔等能够刻画总体单元空间位置的地理信息称作总体单元的空间信息。传统抽样调查方法对于总体单元做出了独立性假设,也即各个单元之间不存在相关关系。如果在抽样调查设计中考虑到总体的空间信息,这样的假定便难以满足。托普勒第一定律指出,任何事物之间都存在相关性,空间距离相近的事物的相似性比距离远的事物相似性更大[1]。这样的例子在抽样调查中不胜枚举,例如:在商业调查中,由于区域经济发达程度的不平衡,营业总额较高的大型商场集中分布在繁华闹市和中心商圈;在社会调查中,居住在不同住宅社区的被访者购买奢侈品的习惯存在较大差异,高档社区的居民具有更强的奢侈品购买力;在自然资源的调查中,某山区的植被种类覆盖情况会受到海拔高度变化的影响。

抽样调查中存在的上述现象被称作空间相关性,也即空间距离相近的总体单元的目标变量存在较强的相关关系,总体单元的规模与其空间位置有关。如果不考虑空间信息的存在,抽样调查的实际工作将面临两个棘手的问题:一方面,样本的随机性不能保证每次抽样都能保证样本的规模结构与总体保持一致,影响样本代表性;另一方面,重复抽样时每次选取的样本结构差异较大,基于样本得出的估计值变异性较大,导致估计效率的下降。因此,在面对空间总体的抽样调查中不能忽视空间相关性的存在,应当利用技术手段刻画总体的空间差异,探索该种情形下提高样本代表性和估计效率的方法。在传统抽样设计中考虑总体的空间属性,主要有如下两种思路:

(1)分层抽样:在传统抽样设计中,以空间区域为分层标志的分层随机抽样在实践中得到了广泛运用。当总体单元的目标变量在不同空间区域之间差异较大时,空间分层抽样能够有效提高样本代

表性和估计效率。但其局限性在于,分层标志往往是诸如行政区划、街道乡镇等定性辅助信息,如果想要对空间差异进行更加精细的刻画,必然要求空间划分的层数越多越好。但是,在实践中寻找到足够精细的分层标志并不容易,即便能够实现足够精细的分层,样本量分配和采集也会由于层数的增多而难以操作。换言之,空间分层抽样面临着可操作性和空间分层精细度的两难选择。

(2)系统抽样:依照空间位置排序后的系统抽样在面对一些特殊的空间总体时非常实用,例如刘蕴芳等(2005)[2]在孝襄高速公路第 9 合同段绿化验收的实际工作中利用一维空间信息构造了等距系统抽样。分层抽样利用的是总体单元的"隶属关系",系统抽样利用的则是总体单元的"相对位置关系"。系统抽样的局限性在于其应用场合较为特殊,不如空间分层抽样适用性强,仅能对公路、航道、河流等单维度下带状分布的空间总体进行调查研究,不适用于二维空间乃至高维空间中的总体。为此,Stevens 等(2004)[3]将系统抽样在二维空间中进行了推广,提出了广义随机棋盘布局抽样(Generalized Random - Tessellation Stratified,简称 GRTS)。GRTS 法利用空间位置信息将总体单元在二维空间中排列,并按照设定的函数投射到一维空间上进行排序,然后使用系统 πPS 抽样对排序后的单元进行抽样,该种方法可被视为系统 πPS 抽样在二维空间中的拓展。但是,该种方法被诸多学者证明存在弊端,因为从二维空间向一维空间投射的过程可能会使原总体的结构改变,因此 GRTS 法可能会导致总体单元在二维空间的相对位置无法在一维空间中准确的表达。由此可见,传统抽样方法在面对空间总体的实际调查工作中仍存在着诸多问题。

本文针对上述问题提出了新的解决思路:充分利用总体中存在的地理坐标信息,借助地理学中的空间平衡抽样方法解决统计学调查中总体空间相关性对样本代表性和估计效率的影响。依照上述思路,本文的讨论从两方面入手:一是在抽样过程中利用空间信息解决空间总体抽样调查中面临的样本代表性问题,二是利用包含在样本中的空间信息改进方差估计量形式,获得估计效率更高的方差估计量。这两方面内容将分别在本文第 1 节和第 2 节分别进行探讨,第 3 节和第 4 节的研究将利用本文提出的抽样方法和估计方法分别进行模拟和实证分析。本文的研究体现着"解铃还须系铃人"的思想,总体单元的空间相关性源于地理坐标信息的纳入,因此在实践中就应当将这部分空间信息在抽样和估计中加以利用,从而解决传统方法面对总体单元空间相关性问题时的困扰。

1 空间平衡抽样

1.1 空间平衡样本的特征

样本的空间平衡性思想起源于对自然资源的调查,自然资源总体具有典型的空间相关性,矿产含量、植被分布以及空气污染等现象都与采样点的空间位置有关。为了在空间总体中获得代表性优良的样本,Stevens 等(2004)[3]最早在对自然资源的调查中提出了通过二维空间中的系统抽样将样本在空间中均匀的排列,从而提高样本代表性和估计效率。那么,在空间中均匀覆盖的样本是否具有空间平衡性呢?

事实上,空间平衡样本是一种特殊的平衡样本,是基于空间信息构造的平衡样本。所谓平衡样本,是指样本依照一系列与目标变量高度相关的辅助变量(也即平衡变量)与总体保持结构上的一致,也即:

$$\sum_{i\in U} x_i = \sum_{i\in S} \frac{x_i}{\pi_i}.$$

上式中,U 和 S 分别表示总体和样本,x 为平衡变量向量,π_i是第 i 个单元的包含概率。通常,样本的平衡性难以严格满足,仅要求$\sum_{i\in U} x_i$与$\sum_{i\in S} \frac{x_i}{\pi_i}$近似相等即可。为了研究平衡样本与空间中均匀覆盖的样本之间的联系,Grafström 等(2013)[4]基于泰森多边形法从数学上证明了对空间形成均匀覆盖的样本能够满足平衡样本的条件。因此,空间平衡样本的重要特征是:样本点能够对空间形成均匀的覆盖。Tille(2006)[5]在研究平衡样本时提出,平衡变量 x 与目标变量的相关程度越强,估计量的估计效率就越高。类似地,空间总体中目标变量与空间位置具有相关性,提高估计效率的途径是利用总体单元的空间信息作为辅助变量,抽取对空间形成均匀覆盖的样本(空间平衡样本)。

1.2 空间平衡抽样算法

空间平衡样本的获取可利用空间平衡抽样设计,最具代表性的两种方法是空间关联泊松算法[6]

(Spatially Correlated Poisson Sampling,简称SCPS)和局部枢轴算法[7](Local Pivotal Methods,简称LPM)。上述两种抽样算法对于任意维度的空间都具有适用性,因此适用于地理坐标信息参与下的二维空间抽样调查。其原理是首先为总体单元设置初始包含概率,初始包含概率向量为 $\pi=(\pi_1,\pi_2,\cdots,\pi_N)$。随后通过算法将各总体单元的包含概率向量更新为入样指示向量 $\pi'=(I_1,I_2,\cdots,I_N)$,从而指示哪些总体单元进入样本。其中:

$$I_i=\begin{cases}1, & \text{若总体单元 } i \text{ 入样},\\ 0, & \text{若总体单元 } i \text{ 未入样}.\end{cases}$$

两种抽样算法均会参考总体单元之间的空间距离,为空间上相邻近的单元赋予高度负相关的包含概率,进而获取对空间具有良好覆盖性的空间平衡样本。SCPS算法和LPM算法有不同的算法结构,通过不同的路径实现了空间平衡样本的获取。

1.2.1 SCPS算法

SCPS法基于总体单元的初始包含概率,对各个单元逐一进行访问,并按照相邻单元之间的空间距离对包含概率进行更新,空间距离相近的单元被赋予高度负相关的包含概率,其算法的详细步骤如下:

(1)为总体单元设置初始包含概率;

(2)依照总体单元间的空间距离构成权重矩阵 $\Delta=(w_j^i)_{N\times N}$,矩阵中元素 w_j^i 表示单元 j 赋予单元 i 的权重,该权重与单元 j 与 i 的空间距离呈负相关;

(3)对所有单元进行序列式访问,每一步对一个单元的初始包含概率使用规则 $\pi_i^j=\pi_i^{j-1}-(I_j-\pi_i^{j-1})w_j^i$ 进行更新,其中 $i>j$,π_i^j 表示单元 i 在第 j 步更新后的包含概率,I_j 表示第 j 个单元的入样指示向量;

(4)基于更新后的包含概率 π_i^j 进行一次随机实现,确定单元 i 的选样结果,若入样则 $I_i=1$,否则为0;

(5)所有单元都被访问后,包含概率向量被更新为入样指示向量,其中只有0和1两个元素,入样指示变量为1的单元构成空间平衡样本。

由SCPS算法的运算步骤可见,第(2)步的计算是该算法的基础,这一步骤度量了总体单元之间相互影响的力度,空间距离越近的单元互相影响的力度被设定的越大,是对空间相关性的定量度量;第(3)步则是整个算法的核心,对第 i 个总体单元的包含概率进行更新时,需参考已被算法完成访问并进行包含概率更新的 j 个单元的入样结果 I_j,单元之间空间距离越近,相互之间的权重影响越大。如此,在依照更新后包含概率进行随机实现时,空间距离相近的单元倾向于不同时进入样本,进而获得对空间形成均匀覆盖的空间平衡样本。

1.2.2 LPM算法

与SCPS算法序列式的更新顺序不同,LPM算法是一种在局部空间中对总体单元进行成对计算的选样算法,其核心思想是直接锁定空间中距离最近的一对单元为其赋予高度负相关的包含概率,通过重复计算访问所有单元后获取的样本具有良好的空间覆盖性,其算法具体步骤为:

(1)为总体单元赋予初始包含概率;

(2)从总体中随机抽取一个单元,计算该单元与其他单元的距离,选取与该单元空间距离最近的单元;

(3)使用如下规则更新抽取的一对单元的包含概率:

$$(\pi_i',\pi_j')=\begin{cases}(0,\pi_i+\pi_j)\text{的概率是}\dfrac{\pi_j}{\pi_i+\pi_j},\\ (\pi_i+\pi_j,0)\text{的概率是}\dfrac{\pi_i}{\pi_i+\pi_j},\end{cases}$$

$\pi_i+\pi_j<1$,

$$(\pi_i',\pi_j')=\begin{cases}(1,\pi_i+\pi_j-1)\text{的概率是}\dfrac{1-\pi_j}{2-\pi_i-\pi_j},\\ (\pi_i+\pi_j-1,1)\text{的概率是}\dfrac{1-\pi_i}{2-\pi_i-\pi_j},\end{cases}$$

$\pi_i+\pi_j\geqslant 1$,

其中,π_i、π_j 为更新运算开始前的包含概率,π_i'、π_j' 为更新后的包含概率。参与运算的一对总体单元的初始包含概率组 (π_i,π_j) 依照 π_i、π_j 的数值大小有可能被算法更新得到四种不同的结果,分别是:$(0,\pi_i+\pi_j)$、$(\pi_i+\pi_j,0)$、$(1,\pi_i+\pi_j-1)$ 和 $(\pi_i+\pi_j-1,1)$,包含概率被更新为0或1的总体单元之后不再参与随机抽样、访问和运算。因此,上文中提到的"更新运算开始前的包含概率"对于未被算法访问的单元而言是指初始包含概率,对于已被算法访问但包含概率未被更新至0或1的单元而言是指上一次运算更新得到的包含概率;

(4)重复(2)和(3)中的成对选样与成对访问更新过程,直到所有总体单元的包含概率被更新为

0 或 1,也即包含概率向量被更新为入样指示向量,入样指示变量为 1 的单元构成空间平衡样本。

由上述步骤可见,LPM 算法的核心是第(3)步,每一步循环计算都将空间中随机选择的一对相邻单元其中之一的包含概率更新为 0 或 1,从而指示其该单元否入样,并依照算法规则为另一个单元赋予与相邻单元高度负相关的包含概率,从而保证了空间距离相近的单元倾向于不同时进入样本,提升样本代表性。另外,在 LPM 算法的第(2)步中,如果要求选取的一对单元互为最邻近单元,则被称为 LPM1 算法,反之被称为 LPM2 算法。LPM1 算法抽取的样本具有更好的空间平衡性,LPM2 算法则拥有更快的计算速度。

1.3 空间平衡抽样的优势和可行性

相较于传统抽样方法中的空间分层抽样,空间平衡抽样算法对空间信息的利用更加精细,不再仅仅局限于利用空间区域分层标志将总体“分类”,而是对总体单元的相对位置作出精确的刻画,总体单元之间的空间距离和相对位置得以展现。相较于传统抽样方法中基于空间位置排序的系统抽样,空间平衡抽样算法不再有维度的限制,在实际工作中的适用性和可行性更强。很多网络地图和移动应用可查询经纬度信息,在抽样框编制阶段可以方便地获取总体单元的地理坐标信息。空间平衡抽样也可以与其他抽样方法结合使用,在政府调查、民意调查和商业调查中常采用多阶段抽样,如果初级抽样单元是区县、机构、商圈、学校等具有实际空间位置的单点,对初级抽样单元抽取空间平衡样本能够有效地提高样本代表性。

2 基于空间平衡样本的估计

空间平衡抽样利用了总体单元的空间信息,因此样本中包含了每个单元的空间位置信息,基于空间平衡样本的统计推断也可以对样本单元的空间信息加以利用,主要体现在借助地理坐标信息改进方差估计量形式进而提高估计效率。

空间平衡样本中的单元对空间具有良好的覆盖性,因此相邻单元之间的空间距离较近。受到空间相关性的影响,距离相邻近的样本单元之间存在相似性,可利用样本中的空间信息刻画样本单元的相对位置,进而对样本单元以若干个中心进行“聚拢”,形成若干个空间区域,将考察单个样本单元的变异性改为考察若干个空间区域之间的变异性。最早利用到空间信息的方差估计量是 Stevens 等(2004)[3] 开发的基于 GRTS 法的最近邻方差估计量:

$$\hat{V}_{NBH}(\hat{Y}) = \sum_{i\in s}\sum_{j\in D_i} w_{ij}\left(\frac{y_j}{\pi_j} - \bar{y}_{D_i}\right)^2,$$

其中 D_i 是由空间中至少四个相邻近单元构成的区域集合,w_{ij} 是与空间距离成反比的权重,且 $\sum w_{ij} = 1$,$\bar{y}_{D_i}$ 是区域 D_i 的目标变量总值。

$\hat{V}_{NBH}(\hat{Y})$ 估计量可被视为区域 D_i 中的元素依照距样本单元 i 的空间距离反比为权重的加权累计总变差,体现了将样本的变异性依照空间位置进行“聚拢”的思想。虽然 GRTS 法本身存在的样本代表性缺陷,最近邻方差估计量在实践中的运用需要承担一定风险,但是该种方差估计量的构造思路可被应用在构造适用于 SCPS 算法和 LPM 算法的空间平衡方差估计 6 量[7]:

$$\hat{V}_{SB}(\hat{Y}) = \sum_{i\in s}\frac{n_i^*}{n_i^* - 1}\left(\frac{y_i}{\pi_i} - \frac{1}{n_i^*}\sum_{j\in s_i^*}\frac{y_i}{\pi_j}\right)^2,$$

其中 s_i^* 是容量为 n_i^* 的样本子集,其中包含样本单元 i 和满足下列条件的样本单元 j:

$$d(i,j) = \min_{k\in s, k\neq i} d(i,k).$$

由于距离单元 i 最近的单元数可能并不唯一,因此 $n_i^* \geqslant 2$。$\hat{V}_{SB}(\hat{Y})$ 估计量利用样本的空间信息,以样本单元为中心计算与其他单元的空间距离,利用 n_i^* 个子集 s_i^* 的变异性代替 n 个单元变异性,对样本随机性带来的变异也即估计量方差进行估计。显然,每一个子集 s_i^* 对于其中包含的全部样本单元之间的变异性进行了“压缩”,估计量的估计效率得以提升。

需要特别指出的是,基于空间平衡样本的总体总值估计量仍可以使用经典的赫尔维兹汤普森估计量(HT Estimator),又称 π 估计量:

$$\hat{Y}HT = \sum_{i=1}^{n} \pi_i^{-1} Y_i.$$

对于 SCPS 和 LPM 算法而言,上式中的包含概率 π_i 为初始包含概率,因为算法将初始包含概率更新为入样指示变量的目的在于选样,每个单元的入样结果都是基于初始包含概率随机实现而得到。因此,空间信息并不参与目标变量的估计,仅在方差估计量中使用。HT 估计量具有无偏性,空间信息在基于空间平衡样本的统计推断中的主要作用

是提高估计效率。

3 模拟研究

为了研究空间平衡样本的效果，现基于二维空间内的模拟总体数据进行研究。总体单元分布在一个 10×10 的二维空间内，总体单元的空间位置坐标可表示为(x,y)，其中 $1 \leqslant x \leqslant 10, 1 \leqslant y \leqslant 10$。总体单元的目标变量数值与其空间位置相关，现构造两个空间总体：

总体 1：$W_1 = x + y$，

总体 2：$W_2 = 3\cos(x) + 2\sin(y) + 5$.

图 1 和图 2 中的黑色圆点大小表示总体单元目标变量的数值，由于总体的目标变量数值与其空间位置存在函数关系，显然两总体的单元之间存在空间相关性。需要特别指出的是，上述两个总体分别刻画了不同的总体特征。由图 1 可见总体 1 中的目标变量 W_1 随着坐标变化而单调变化，在 x 轴和 y 轴两个方向上，目标变量数值都随着坐标点数值的增大而增大，目标变量数值较大和较小的单元分别集中分布在总体空间的右上角和左下角；由图 2 可见，总体 2 中的目标变量 W_2 在 x 轴和 y 轴两个方向上均不是单调变化，总体空间中存在若干个目标变量数值较大和较小单元的集中分布区域。下文的研究将进行多次重复抽样和估计，以对比不同特征的总体中不同样本量下空间平衡抽样和简单随机抽样在样本代表性方面的表现，并对比空间平衡抽样下不同方差估计方法的估计效率。

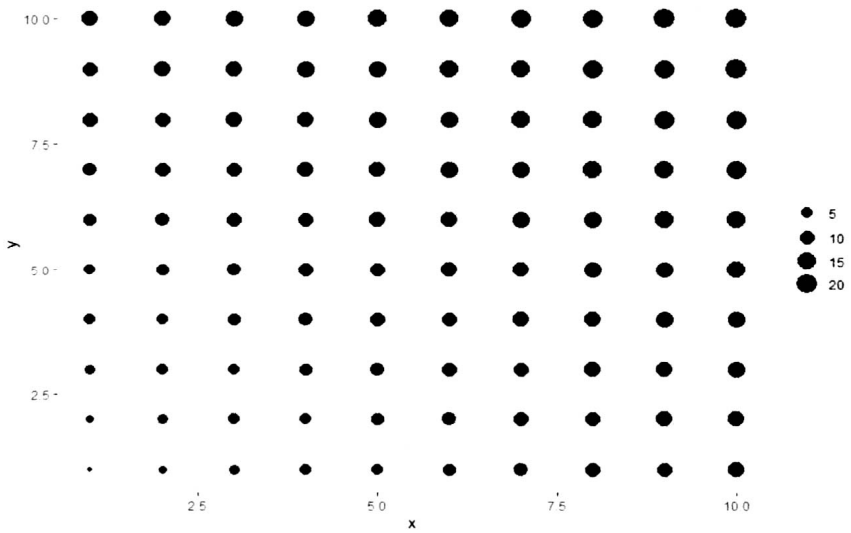

图 1　总体 1 的空间相关性

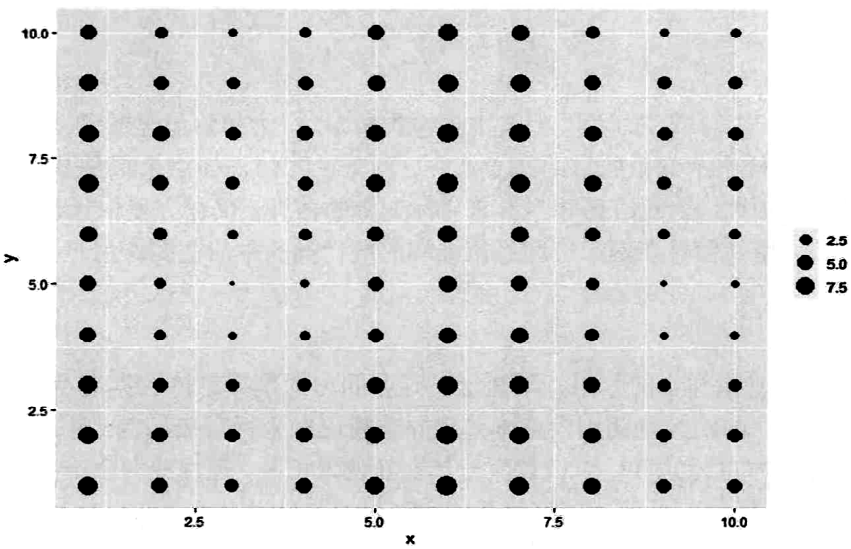

图 2　总体 2 的空间相关性

现使用 R 软件中的 BalancedSampling 程序包对上述总体进行等概率抽样情形下的抽样，利用简单随机抽样、SCPS 算法和 LPM 算法分别抽取容量 n = 50 和 n = 20 的样本各 1000 次，汇总基于 1000 个重复样本的总体总值估计结果均值汇总如表 1。

表 1　各种抽样方法 1000 次重复抽样总体总值估计结果对比

抽样方法	待估指标	模拟总体 1		模拟总体 2	
		$n=50$	$n=20$	$n=50$	$n=20$
简单随机抽样	$\hat{W}$	1100.06	1100.69	484.63	484.96
	$\hat{\sigma}_{sim}(\hat{W})$	40.70	80.84	25.04	49.59
SCPS 算法	$\hat{W}$	1099.76	1099.59	485.98	486.13
	$\hat{\sigma}_{sim}(\hat{W})$	41.05	83.19	25.17	50.68
	$\hat{\sigma}_{SB}(\hat{W})$	9.20	25.09	16.44	43.13
LPM 算法	$\hat{W}$	1099.91	1098.99	485.89	485.89
	$\hat{\sigma}_{sim}(\hat{W})$	41.04	83.08	25.19	50.60
	$\hat{\sigma}_{SB}(\hat{W})$	8.76	25.13	16.05	41.91

总体 1 的总体总值真值为 1100，总体 2 的总体总值真值为 485.70。由表 1 可见，在方差估计中，如果不考虑样本的空间特性，使用简单随机抽样的方差估计量 $\hat{V}_{sim}(\hat{W})$，相同样本量下标准差的估计值差别并不大。但是，空间平衡方差估计量 $\hat{V}_{SB}(\hat{W})$ 的估计效率较 $\hat{V}_{sim}(\hat{W})$ 有显著的提升，这是由于 $\hat{V}_{SB}(\hat{W})$ 估计量使用空间信息对估计量的形式进行了优化改进，使估计效率得到了额外的提升。尤其是在小样本量情况下，空间平衡方差估计量能够很大程度上的提高估计效率，在实际工作中能够降低对样本量的要求。另外，无论对于大样本还是小样本，空间平衡方差估计量对估计效率的提升在总体 1 要比总体 2 中更为显著，这是由两个模拟总体不同的特征决定的。

上述模拟研究中使用的总体总值估计量是经典的 HT 无偏估计量，因此几种方法在理论上均可得到总体总值的无偏估计。但是对于一次抽样估计来说，偏差仍然是存在的，现使用均方误差指标进一步对比不同抽样和估计方法的估计效率。在计算均方误差时需要使用方差估计量，其中简单随机抽样采用简单方差估计量，两种空间抽样方法均采用空间平衡方差估计量，汇总 1000 次抽样估计的均方误差均值如表 2 所示。

表 2　各种抽样方法 1000 次重复抽样均方误差均值对比

抽样方法	模拟总体 1		模拟总体 2	
	$n=50$	$n=20$	$n=50$	$n=20$
简单随机抽样	3326.19	12813.06	1293.97	5014.09
SCPS 算法	159.05	1112.79	418.50	2722.98
LPM 算法	186.69	1417.48	442.31	2793.92

由表 2 汇总的结果可见，两种空间抽样算法和空间平衡方差估计量能够显著提高估计效率。现进一步考察两种空间抽样算法对于样本代表性的提升，不妨以总体 2 为例，使用三种抽样方法分别进行一次样本量为 20 的抽样，样本点在总体空间中分布的对比如图 3。

由图 3 到图 5 可见，LPM 算法获取的样本具有最好的空间覆盖性，SCPS 算法获取的样本覆盖性与 LPM 算法较为接近，但简单随机抽样获取的样本的空间覆盖性则明显逊色于空间平衡抽样方法。在总体中存在空间相关性的情况下，空间中距离相近的样本具有更多的相似性，空间平衡抽样将样本点在空间中形成均匀的覆盖，从而提升了样本代表性。

综上，空间信息参与抽样设计和估计时，空间平衡抽样能够消除空间相关性的不利影响，获得代表性更强的样本，并基于空间平衡样本构造统计性质更加优良的估计量。SCPS 与 LPM 两种算法在实际工作中的操作方法并无区别，具有相同的适用性。上述结果是基于对模拟总体的抽样而得出的，总体单元之间具有函数形式的完全空间相关性，也即总体单元的目标变量数值完全由空间位置决定。前文中提到，平衡变量与目标变量之间相关性越强，基于平衡样本的估计效率就越高。在抽样调查的实际工作中，与总体目标变量相关的混杂因素较多，总体规模与空间位置的相关关系并非严格的函数关系。因此，在模拟研究证实空间平衡样本存在样本代表性和估计效率优势的基础上，有必要结合实际案例对空间平衡样本的应用进行探讨。

4　实证研究

本文以北京市 251 家非分支社区卫生服务中心在岗职工总数抽样调查为例进行研究。

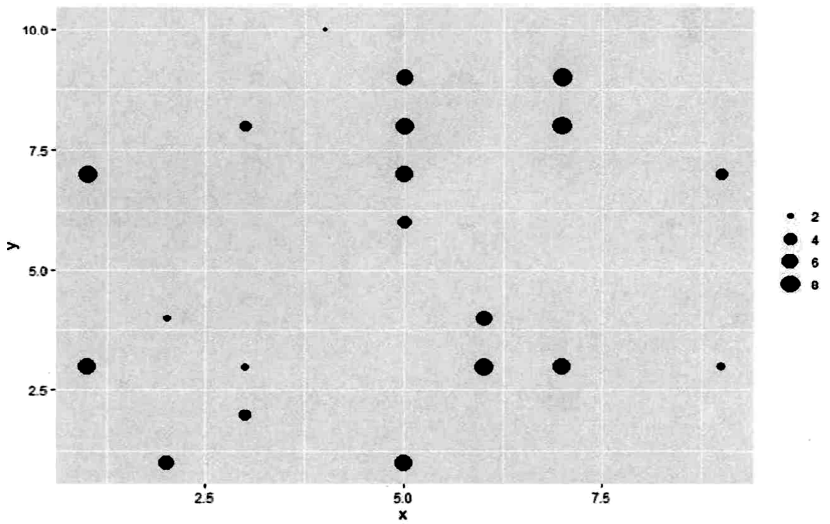

图 3　简单随机样本的分布

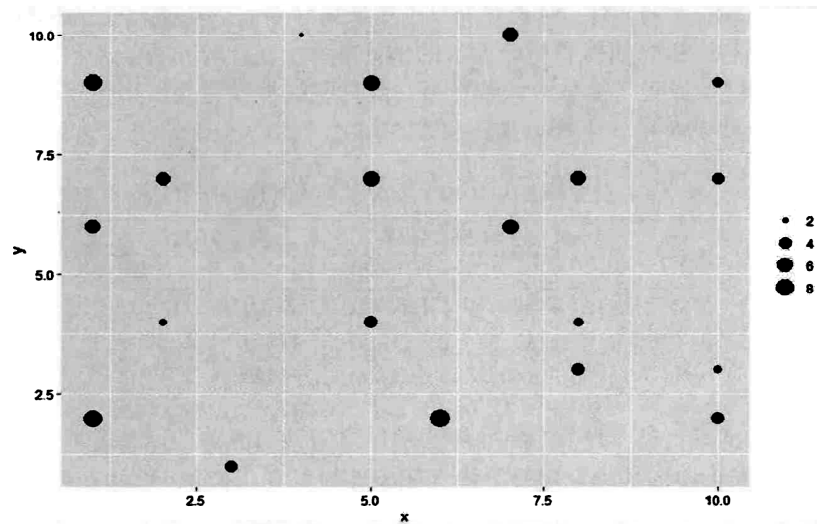

图 4　SCPS 空间平衡样本的分布

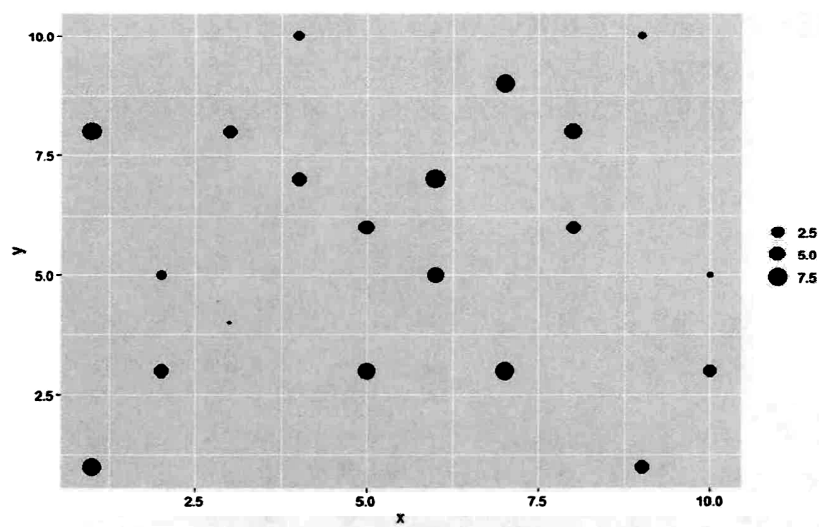

图 5　LPM 空间平衡样本的分布

4.1 案例背景

自北京与2017年4月8日施行医改政策以来，分级医疗的大趋势已经建立，小病上社区的理念在人民群众当中逐渐树立，以医疗资源合理配置为目的政策效果逐渐凸显。在医改逐渐深入的政策背景下，对于医疗资源的定期调查能够为医疗行政部门提供有力的数据支撑，为政策的推行提供参考，具有很强的实际意义。由于医疗资源的分布具有空间相关性，因此在该种类型的调查中可考虑使用空间平衡样本进行推断。本案例采用北京市251家非分支社区卫生服务中心作为研究总体，目标变量为总体中251家非分支社区卫生服务中心（以下简称“中心”）的在岗职工总数。

4.2 数据来源

本案例中抽样框名录来源于北京市卫生信息汇编材料，用于构造空间平衡样本的空间信息为各家中心的经纬度坐标。其中，各家中心的在岗职工数据可由北京市卫生信息汇编材料中查询获得，经纬度信息可利用网络地图的坐标拾取功能采集。本案例基于真实总体进行实证研究，方便对空间平衡样本在实际工作中的效果进行评价。

4.3 总体的空间特征

为了展示总体单元的空间相关性，绘制气泡图如图6所示，图中横坐标为东经坐标，纵坐标为北纬坐标，气泡大小代表总体单元的目标变量值大小，也即在岗职工数的规模。

由图6可见，规模较大的医疗机构集中分布在中心城区。由于气泡有重叠，规模较小的医疗机构分布情况从图6中难以进一步识别，现截取东经116.3至116.4度，北纬39.8度至40.1度的部分单

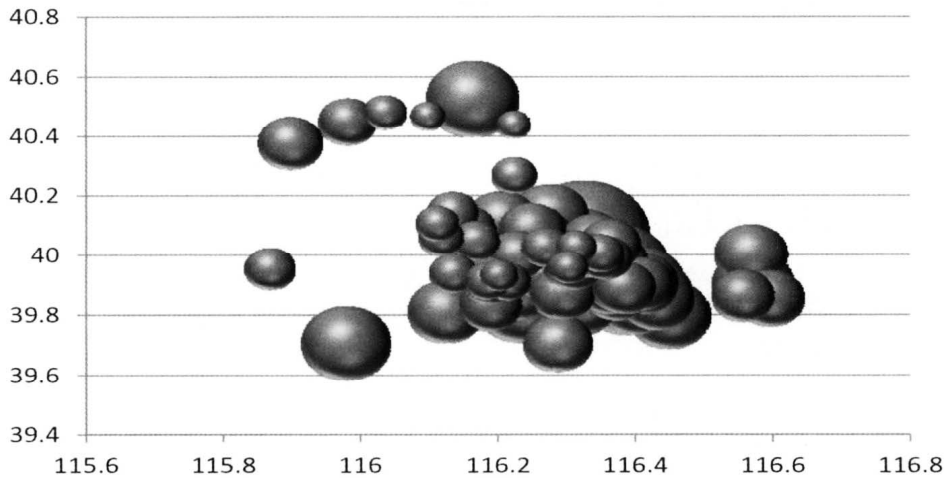

图6 总体单元的空间相关性

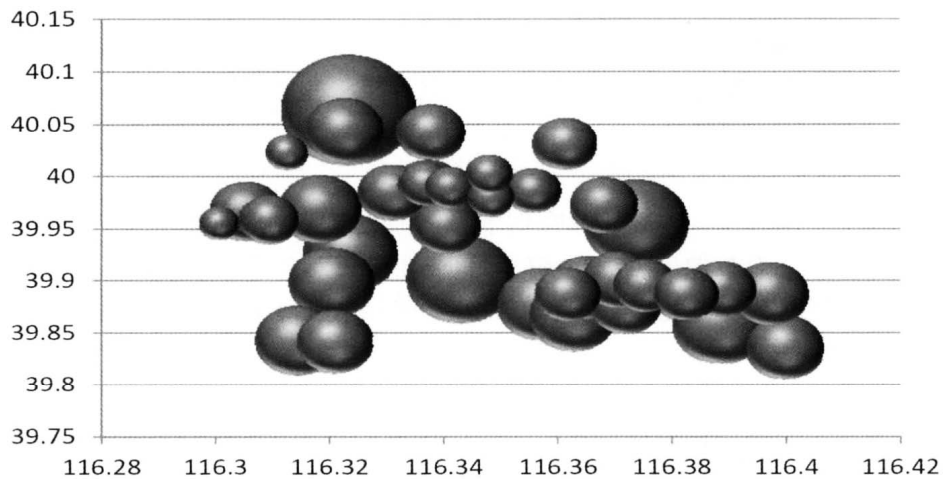

图7 总体局部的空间相关性

元,将图 6 局部放大,如图 7 所示,进一步讨论总体的空间相关性。

由图 7 可见,在局部空间中,规模较小的单元呈现出了集中分布的趋势,在北纬 39.95 至 40 度的区域内,有近似于一字排开的一系列中小规模的单元;在北纬 39.85 至 39.9 度、东经 116.38 度附近的区域内,也有一系列中小规模的总体单元呈现出了一字排开的趋势。另外,规模相对较大的单元在图 7 的局部空间中分布相对分散,但是在图 6 的总体空间下仍具有一定的聚集特征。

4.4 抽样设计

为了验证空间平衡样本在实际案例中的效果,不妨在本案例的研究过程中同时抽取简单随机样本、分层随机样本和空间平衡样本进行对比。简单随机样本的获取可直接利用抽样框名录进行随机抽取;分层随机样本以城区和郊区作为分层标志,层内样本量采用比例分配;空间平衡样本的获取采用 SCPS 算法和 LPM1 算法进行,两种算法样本选取过程基于反复计算空间距离并更新包含概率而进行,具体算法步骤可参考前文。实际操作中,使用 R 软件中的 BalancedSampling 程序包选取空间平衡样本。使用 R 软件编写程序对总体进行等概率重复抽样,依照 20% 的抽样比抽取容量为 50 的简单随机样本、SCPS 空间平衡样本和 LPM1 空间平衡样本各 1000 次。

4.5 基于样本的估计

由于本文具备实际的总体数据,可以用来对不同抽样方法获得的样本的估计结果进行对比评估,因此,直接利用 R 软件选样结果进行计算便可完成估计,1000 次估计的估计量均值和标准差情况可见表 3 汇总,其中在岗职工人总数估计采用 HT 估计量,空间平衡样本的方差估计采用空间平衡方差估计量。表中抽样误差计算公式为 $1.96\times\frac{\hat{\sigma}}{\sqrt{n}}$,相对误差计算公式为 $1.96\times\frac{\hat{\sigma}}{\sqrt{n}}\div\hat{Y}_{HT}\times100\%$。

4.6 讨论

由表 3 可见,空间平衡样本在实际案例中的估计效率优于传统抽样方法获取的样本,能够对各中心的在岗职工总数做出更加精确的区间估计,相对抽样误差有所下降。在实际案例的研究中,基于空间平衡样本的推断提升估计效率的程度较模拟研究的结果有所下降,这是由于总体的目标变量还会受到其他混杂因素的影响。总体单元规模差异受空间位置影响的程度越大,空间平衡样本提升估计效率的程度也随之增大。由此可见,使用 SCPS 和 LPM 算法获取的空间平衡样本虽然能保证样本代表性,但是在提升估计效率方面的表现取决于总体中空间相关性的强弱。当总体中存在很强的空间相关性时,SCPS 和 LPM 算法获取的空间平衡样本能够显著的提高估计效率;若总体中存在其他强烈影响总体规模的混杂变量,空间位置并非唯一影响总体规模的因素时,总体在局部可能呈现出一定的异质性,使用基于 SCPS 和 LPM 算法获取的空间平衡样本进行推断的优势会受到一定损失。考虑模拟研究结果,空间平衡抽样算法和空间平衡方差估计量的优势在局部异质性更大的总体 2 中不如在均质性更强的总体 1 中明显;再以本文图 7 展示的实证研究局部总体特征为例,在局部空间中虽然规模相近的总体单元在空间中聚集分布,但在一个相对较小的空间区域内总体单元也存在一定的异质性。该现象说明,总体单元之间的差异可能受到其他因素的影响,在本案例中社区医疗机构的在岗职工人数有可能受到社区规模、区域经济情况以及社区

表 3　　1000 次重复抽样总体总值估计结果对比

样本类型	总体总值估计值	标准差估计值	抽样误差	相对误差(%)
简单随机样本	25547.40	2342.12	649.20	2.54
分层随机样本	25392.83	2143.39	594.12	2.34
SCPS 空间平衡样本	25448.82	1931.95	535.51	2.10
LPM 空间平衡样本	25482.48	1950.42	540.63	2.12

人口数等因素的影响。但即便如此,地理坐标信息参与空间平衡抽样和方差估计时,样本的代表性和估计效率均优于传统抽样和估计方法,只是改进的程度取决于总体的局部异质性以及有无其他与总体规模相关的混杂因素。

5 结论

地理信息技术的发展为抽样调查的实际工作中提供了可利用的地理坐标信息,构造抽样框阶段可采集总体单元的空间位置信息并加以利用,进而抽取空间平衡样本。基地理坐标参与抽样和估计带来两方面优势:一是空间平衡抽样解决了空间相关性对样本代表性的影响,二是基于空间信息构造的方差估计量能够在空间相关性较强的情形下提高估计效率。

在利用地理坐标信息进行空间平衡抽样和估计时,如果存在其他造成总体规模差异的混杂因素,可考虑两种解决思路:一是在构建空间平衡样本的过程中,将空间信息与传统的辅助变量同时加以利用,例如 Grafström 等(2013)[8]提出在空间抽样中加入其他平衡变量,构造空间双重平衡样本;二是在估计阶段引入模型或随机过程,例如李政寰等(2018)[9]基于最大稳定过程使用最大熵法选择北京市空气质量监测点的位置。上述两种思路为解决前文所述的问题提供了思路,有待学者们结合实际展开更加深入的研究。

参考文献:

[1] Tobler W A. Computer movie simulating urban growth in the Detroit region[J]. Economic Geography,1970,46(2):234 - 240.

[2]刘蕴芳,王慧觉. 高速公路绿化工程验收中的抽样方法研究[J]. 武汉理工大学学报(交通科学与工程),2005,29(6):1001 - 1004.

[3] Stevens D, Olsen A. Spatially balanced sampling of natural resources[J]. Journal of the American Statistical Association,2004,99(465):262 - 277.

[4] Grafström A, Lundström N. Why well spread probability samples are balanced[J]. Open Journal of Statistics,2013,3(1):36 - 41.

[5] Tille Y. Sampling Algorithm[M]. New York: Springer Science + Business Media, Inc.,2006.

[6] Grafström A. Spatially correlated Poisson sampling[J]. Journal of Statistical Planning and Inference,2012,142(1):139 - 147.

[7] Grafström A, Lundström N L P, Schelin L. Spatially balanced sampling through the pivotal method[J]. Biometrics,2012,68(2):514 - 520.

[8] Grafström A, Tille Y. Doubly spatial sampling with spreading and restitution of auxiliary totals[J]. Environmentrics,2013,24(2):120 - 131.

[9]李政寰,金勇进. 基于最大稳定过程的北京市 PM2.5 监测站空间网络设计[J]. 数理统计与管理,2018,37(5):871 - 879.

Spatial Balanced Sampling Design with the Participation of Geographic Coordinate Information

Hao Yiwei Jin Yongjin

Abstract: Spatial dependency in surveys breaks the assumption of independence in traditional sampling methods. Spatial balanced sample is used to improve the representative of sample and spatial auxiliary information is used to improve the variance estimator in order to gain more efficiency. It is shown by simulation and case study that statistical inference based on spatial balanced sample gains more efficiency of variance estimator under a strong spatial dependency in the population.

Key words: geographic coordinate information; spatial information; spatial balanced sample; spatial dependency