空间人口数据间接估计方法与应用研究

王广州

【提要】 本文通过对空间克立格算法和人口应用实例的讨论,提出在人口数据间接估计过程中改进估计过程的方法。

【作者】 王广州 北京大学人口研究所,博士研究生。

空间信息是描述事物属性的重要信息之一。虽然空间统计方法始于50年代,然而,由于受数据收集等因素的影响和限制,空间统计方法并没有得到应有的重视和广泛应用。近年来随着计算机技术的迅猛发展,尤其是地理信息系统方面的研究日益加强,国外在空间统计分析方法的研究和应用方面得到异常迅速发展(Brian D.,1981)。与国外相比,国内在这方面的研究和应用相对比较落后,仅仅是80年代以来,才受到生态学和地质学等领域学者的高度重视(王政权,1999;周国法等,1998;於崇文等,1980)。

对于人口学者来说,间接估计已成为人口学分析方法的重要组成部分(曾毅,1993)。众所周知,人口学的间接估计方法是采用不同来源人口数据进行间接估计的,如用曾生子女存活比估算婴儿死亡率等。其估计过程是利用人口群体本身的内在联系进行间接估计。然而,在人口统计分析领域一直比较缺乏空间统计分析方面的研究。基于对人口分布和空间特征基本规律的认识,本文试图在人口数据的空间间接估计及其进一步改进方面进行尝试,以期起到抛砖引玉的作用。

1. 方法选择

空间克立格方法最早由南非矿山工程师克立格(D. G. Krege)和统计学家西舍尔(H. S. Sichel)提出来的。法国统计学家G. Matheron在克立格和西舍尔的研究基础上进行了大量的理论和实际研究,在很大程度上推动了空间克立格方法的研究和发展。此后的一些研究不仅对经典空间克立格方法进行大量的理论推导和证明,而且使该方法日益丰富和发展(王政权,1992;於崇文,1980)。目前空间克立格方法已成为空间统计学最重要的理论和方法之一。根据人口问题的特点,本文选用比较常用的普通克立格方法中的对不规则网格取样方法进行人口属性的空间间接估计。其具体实现方法和过程如下:

1.1 数据输入

数据输入包括底图数字化和属性数据库建立两方面工作。数字化过程就是将所研究区域地图底图和样本点相对坐标输入计算机的过程,也就是对整个待研究地区的空间结构数字化。本文采用计算机扫描方式将地图底图输入计算机,然后,随机选定待估样本点并对样本点空间位置数字化。数字化完毕后,建立相应样本点的人口属性数据库。

1.2 算法

假定用空间点 $Z_1(X_1,Y_1,P_1)$,$Z_2(X_2,Y_2,P_2)$,……$Z_n(X_n,Y_n,P_n)$的数据对未知点 $Z_0(X_0,Y_0,P_0)$进行间接估计。X_i、Y_i 为 Z_i 点的空间坐标,P_i 为 Z_i 点的人口属性。

(1) 求各点间距离 D_{ij}:$D_{ij}=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2}$;式中 $i,j=0\cdots\cdots n$。

(2) 选择空间变异函数模型:假定所研究人口指标的空间特征满足二阶平稳和本征假设,故选定指数模型进行空间估计。

(3) 计算指数模型空间变异协方差函数矩阵:

计算 $C_{ij}=\begin{cases}10 & D_{ij}=0\\ 10exp(\dfrac{-3\cdot D_{ij}}{10}) & D_{ij}>0\end{cases}$;得到矩阵 $\begin{bmatrix}C_{11} & C_{12} & \cdots & C_{1n}\\ C_{21} & C_{22} & \cdots & C_{2n}\\ \cdots & \cdots & \cdots & \cdots\\ C_{n1} & C_{n2} & \cdots & C_{nn}\end{bmatrix}$ 和 $\begin{bmatrix}C_{01}\\ C_{02}\\ \cdots\\ C_{0n}\end{bmatrix}$。

(4) 计算克立格权重系数:

计算 $\lambda=\begin{bmatrix}\lambda_1\\ \lambda_2\\ \cdots\\ \lambda_n\\ \mu\end{bmatrix}=\begin{bmatrix}C_{11} & C_{12} & \cdots & C_{1n} & 1\\ C_{21} & C_{22} & \cdots & C_{2n} & 1\\ \cdots & \cdots & \cdots & \cdots & 1\\ C_{n1} & C_{n2} & \cdots & C_{nn} & 1\\ 1 & 1 & 1 & 1 & 0\end{bmatrix}^{-1}\cdot\begin{bmatrix}C_{01}\\ C_{02}\\ \cdots\\ C_{0n}\\ 1\end{bmatrix}$;其中 $\sum_{i=1}^{n}\lambda_i=1$。

(5) 求未知点的属性估计值:计算 $Z_0(P_0)=\sum_{i=1}^{n}(\lambda_i\cdot Z_i(P_i))$。

(6) 计算未知点的克立格估计方差:计算 $\sigma^2=\overline{C}_{vv}-\sum_{i=1}^{n}(\lambda_i\cdot C_{0n})+\mu$。

1.3 方法的改进

从克立格方法本身出发,提高估计精度的方法是正确选择空间变异函数,使之更符合客观实际。若从空间克立格方法的估计过程和对问题性质的研究来考虑,由于空间克立格方法是建立在变异函数理论和结构分析的基础上,它是在有限区域内对区域化变量的取值进行无偏最优估计的一种方法。因此,笔者认为在不改变空间变异函数的前提下,首先对样本点进行筛选,然后进行估计,其结果必定可以提高估计的精度。筛选的过程就是试图排除不属于或有可能与空间变异函数不贴切的部分,并找到选点的最佳方向。也就是说,先利用空间已知点和待估点的已知属性进行聚类分析,然后对未知点的未知属性进行空间克立格分析,最终达到提高对未知点未知属性估算精度的目的。

表 1 空间人口间接估计样本点原始数据表

地区	出生率(‰)	死亡率(‰)	文盲率(%)	地区	出生率(‰)	死亡率(‰)	文盲率(%)
依安县	22.43	4.74	21.01	肇东市	21.59	4.96	19.37
克山县	18.06	4.51	15.99	海伦市	18.72	5.13	20.93
克东县	18.84	5.16	20.72	望奎县	18.61	5.10	16.73
拜泉县	21.50	5.20	21.66	兰西县	20.84	4.93	18.46
大庆市	17.25	2.97	11.15	青冈县	19.86	5.17	16.38
绥化市	17.67	5.08	16.78	明水县	21.81	4.78	19.24
安达市	19.20	5.11	14.34	绥棱县	18.78	5.34	17.31

2. 应用实例

现以中国第四次人口普查黑龙江省数据为人口属性数据，进行人口属性的间接估计和评价。以《中华人民共和国分省地图集》为底图进行空间距离的计算。同时，假定每个区域的行政中心所在地的空间位置标识所属区域的人口属性。所研究区域的人口属性数据和空间分布状况分别见表 1 和图。

图　空间人口指标间接估计样本点扫描图

2.1　样本点的选取

任意选取拜泉县、望奎县、兰西县和安达市作为待估样本，用待估样本周围的点作为空间克立格估计的背景值。以就近为原则进行样本点的选取。故选定依安县、克东县、克山县、明水县、海伦市、绥棱县的人口数据对拜泉县进行估计。选用明水县、海伦市、绥棱县、青冈县、绥化市的人口数据对望奎县的相应人口指标进行估计；选用肇东市、青冈县、绥化市、明水县、绥棱县的人口数据估计兰西县的相应人口指标；选用明水县、青冈县、肇东市、大庆市的人口数据估计安达市的相应人口指标，得到拜泉县、望奎县、兰西县和安达市 4 区域的估计结果(见表 2)。

从上述估计数据与人口普查数据的比较可以看到，出生率和死亡率的相对误差较小，估计结果比较理想；文盲率的相对误差较大，其原因是实际的空间变异函数与采用的指数模型的差别较大。

表 2　样本点人口指标间接估计数据

	拜泉县		望奎县		兰西县		安达市	
	估计值	相对误差	估计值	相对误差	估计值	相对误差	估计值	相对误差
出生率	19.775	0.080	19.369	0.041	19.943	0.0430	20.106	0.0472
死亡率	4.943	0.049	5.100	0.0001	5.066	0.0275	4.46	0.127
文盲率	19.200	0.114	18.125	0.083	17.817	0.035	16.494	0.15
克立格估计值方差	8.330		7.996		8.000		7.500	

2.2　估计过程的改进

为了使估计值更准确，剔除与空间变异函数或各点中不一致指标的影响，笔者建议首先进行待估点周围各点的聚类分析。聚类分析可以使用容易获得的指标进行，然后根据聚类结果剔

除不属于同类的点，最后仍用空间克立格方法对待估点进行估计。本例假定已获得所有样本点的出生率和死亡率数据，并对待估点文盲率进行间接估计。以出生率和死亡率为聚类指标，采用动态聚类法对所有样本点进行聚类，结果如下：

一类：依安县、拜泉县、肇东市、兰西县、青冈县、明水县；二类：克山县、克东县、绥化市、安达市、海伦市、望奎县、绥棱县、大庆市。

表 2 中对拜泉县的估计是根据拜泉周围且距拜泉较近的点进行估计的。通过上述聚类可知，克东县、克山县、海伦市、绥棱县与拜泉县并不属于一类，因此剔除后对拜泉县重新进行估计。对望奎县、兰西县和安达市的重新估计也采用相同的方法进行剔除。由于对安达市的估计在上述所选点中不一致点剔除后，仅剩大庆市一个点。为了提高估计的可靠性，在安达市的周围，选择了与剔除点有较小屏蔽关系的绥化市与大庆市共同进行估计。重新估计文盲率的具体结果见表 3。

表 3　聚类后样本点人口指标间接估计数据

	拜泉县		望奎县		兰西县		安达市	
	估计值	相对误差	估计值	相对误差	估计值	相对误差	估计值	相对误差
文盲率	20.125	0.071	18.337	0.096	17.876	0.032	13.944	0.0276
克立格估计值方差	5.000		6.656		5.00		5.000	

对比表 2 和表 3 可以看到，表 3 对文盲率的估计精度有了进一步的提高。估计精度之所以有了提高，其原因在于通过聚类分析后，选取空间点中空间变异规律更加一致的点参与估计。或者说，在空间上尽量使用那些空间变化规律一致的点，并将不一致的点予以剔除。

本文仅以指数空间变异函数模型为例探讨了人口数据的空间间接估计方法。指数模型仅适用于中等连续性的空间数据，而对于其他空间变异函数模型是否更适合人口数据的空间间接估计，或哪一种空间变异函数模型更适用于不同人口属性指标的空间间接估计还有待于今后的深入研究。

参 考 文 献

1. 曾　毅：《人口分析方法与应用》，北京大学出版社，1993 年。
2. 王政权：《地统计学及在生态学中的应用》，科学出版社，1999 年。
3. 周国法、徐汝梅：《生物地理统计学——生物种群时空分析的方法及其应用》，科学出版社，1998 年。
4. 於崇文等：《数学地质学的方法与应用》，冶金工业出版社，1980 年。
5. 《中华人民共和国分省地图集》，中国地图出版社，1995 年。
6. Brian D. Ripley, Spatial Statistics, John Wiley & Sons 1981.

（本文责任编辑：　朱　萍）