年龄别生育率与总和生育率间接估计方法与应用研究

王广州

【提要】 本文讨论了人口分析技术与遗传算法相结合的年龄别生育率与总和生育率间接估计方法，并以人口普查数据为例展示年龄别生育率与总和生育率间接估计方法的建模过程与实施步骤，同时指出今后有待深入研究的问题。

【作者】 王广州　哈尔滨工业大学管理学院，在站博士后。

年龄别生育率与总和生育率是人口研究的重要指标。年龄别生育率不仅可以从生育模式的角度反映育龄妇女的生育过程，而且可以从生育水平的角度反映育龄妇女的生育状况。按研究问题角度不同对生育水平的测度可以分为队列和时期两种方法。按队列分析育龄妇女的生育水平是对特定队列妇女生育史的回顾，通过队列年龄别生育率可以得到相应队列的终身生育率或队列总和生育率。队列年龄别生育率和队列终身生育率较好地反映了妇女的生育史。按时期指标分析育龄妇女的生育水平则是对特定时期妇女生育状况的研究。时期年龄别生育率与总和生育率从生育模式和生育水平两个方面来反映特定时期育龄妇女的生育状况，是以假想队列为前提的。因此，在反映生育模式和生育水平变化的过程中离不开对年龄别生育率和总和生育率的评判，尤其是反映时期综合生育指标——总和生育率已成为人口与计划生育研究的指示性指标，在人口与计划生育研究过程中起着举足轻重的作用。无论是从应用人口分析的研究角度看，还是从人口分析数学方法研究的角度看，对时期年龄别生育率和总和生育率的研究一直都是人口研究工作的重点，研究文献不胜枚举。特别是1998年邦加茨和费尼《生育的数量与进度》一文发表以来(Bongaarts，J.，et al.，1998)，从方法论角度再次引起了人们对生育率研究的高度重视和广泛的兴趣(郭志刚，2000)。然而，纵观生育率研究方法可以看到，目前对生育率的研究往往需要大量的数据，缺乏基于少量高质量数据前提下对时期年龄别生育率和总和生育率变化过程进行间接估计方法的研究。对时期年龄别生育率与总和生育率变化过程的间接估计不仅有利于深入分析人口系统的生育行为与过程，而且有利于对现有数据进行深入挖掘。本文试图将人口分析技术与现代遗传算法相结合对时期年龄别生育率和总和生育率状况及其变化过程进行间接估计，以期为相关研究提供参考。

一、研究方法

从现有研究文献看，时期总和生育率可以通过一般生育率(GFR)或粗出生率(CBR)与总和生育率的近似关系，即：$TFR \approx 30 * GFR \approx 30 * 4\frac{1}{2} * CBR$ 等方法对总和生育率进行简捷的近似计算(David P. Smith，1992)，但由于其假设条件太强或近似程度相对较差，因此在生育研究过程中受到一定的限制。可是，生育水平经常成为研究者和政策制定者关注与争论的焦点，其根本原因不仅与生育率在人口系统中的重要地位密切相关，而且还与人口调查数据质量密切相关。正是由于该原因，经常发生对生育水平评判的争论。此外，从研究的角度看，在对生育水平的争论过程中，生育

水平研究方法的相对欠缺应该负有一定的责任。因此除了直接调查外，有效的间接估计方法就显得非常重要。与对总和生育率的间接估计或近似计算不同的是，虽然也有通过标准生育表方法对年龄别生育率进行间接估计的方法，但由于对时期年龄别生育率的间接估计必须采用标准生育模式和被估计时点部分年龄组年龄别生育率数据，因此现有方法在应用上受到限制(Samuel H. Preston, et al.,2001)。此外，还可以运用育龄妇女所生子女等方法对年龄别生育率进行间接估计，但由于所需数据是以个体记录为研究单位，需要进行大规模的数据匹配，在具体运用过程中由于原始数据的不易获得而受到限制。如果降低对高质量数据的依赖程度，以一次调查汇总数据为基础对年龄别生育率及其变化过程的研究，由于使用数据量较小，常规估计方法很难得到满意和可靠的结果，所以探索新途径进行有效的间接估计显得非常重要。正是由于使用一次调查数据对年龄别生育率间接估计方法难度相对较大和过于复杂等原因，到目前为止，尚未见相关研究成果的发表。

本项研究试图以一次调查数据(如普查数据)将人口分析技术与现代遗传优化建模方法相结合对年龄别生育率与总和生育率进行间接估计。具体对时期年龄别生育率和总和生育率的间接估计方法可以划分为倒推育龄妇女年龄结构和出生人口数(通过人口回溯获得理论出生人口数)、出生人口预测和遗传算法优化三部分。在这三部分中人口回溯模型解决各年度育龄妇女年龄结构和理论出生人口数的间接估计问题；通过遗传算法得到年龄别生育率与育龄妇女年龄结构相结合可以建立年龄别生育率与模型预测出生人口数之间的相互关系；将理论出生人口数与模型预测出生人口相结合得到遗传算法评估函数，解决遗传算法的目标优化问题，通过评估函数筛选得到最优年龄别生育率。

(一) 年龄别育龄妇女人口数与出生人口数间接估计

对育龄妇女年龄结构和出生人口数的估计可以利用一次人口调查数据对以往人口年龄结构进行回溯，也就是采取“存活倒推法”进行。该方法是根据年龄别人口存活概率和现有人口结构对历史人口进行推测，通过回溯方法估算各年份年龄别人口数。具体算法为：${}_nP^{t_1}_{(x)} = {}_nP^{t_2}_{(x+n)} * (\frac{{}_nL_{(x)}}{{}_nL_{(x+n)}})$；式中${}_nP^{t_1}_{(x)}$是在 t_1 时刻年龄在 x 岁至 x+n 岁的人口数；${}_nP^{t_2}_{(x+n)}$是在 t_2 时刻年龄在 x+n 岁至 x+2n 岁的人口数；${}_nL_{(x)}$为确切年龄在 x 至 x+n 队列存活人年数；${}_nL_{(x+n)}$为确切年龄在 x+n 至 x+2n 队列存活人年数(王广州,2000)。

(二) 基于遗传算法年龄别生育率与总和生育率间接估计优化模型

1. 遗传算法

遗传算法是以生物进化机制为基础，通过计算机数值模拟方法来构造人工系统仿真模型，是计算机科学与生物进化理论相结合的产物。早在20世纪50年代和60年代，计算机科学家开始尝试将生物进化理论与思想引入计算机优化计算。20世纪60年代中期，美国 Michigan 大学的 John Holland 在前人研究工作的基础上提出了位串编码技术，该技术完全适用于生物进化的变异操作和交叉(杂交)操作。正是由于 Holland 提出通用编码技术和简单有效的遗传操作为遗传算法的广泛成功应用奠定了基础。经过近30年的不懈努力，遗传算法不仅已经发展成为解决复杂全局优化问题的重要方法，而且在很多研究领域的具体科学实践中得到广泛的应用并获得成功。

从图1可以看到，遗传算法包含以下几个基本步骤：(1) 选择问题的编码方式，对初始种群的 N 个染色体进行初始化，初始化的方法、染色体的编码方式和取值范围根据具体问题的特征而定，本文采用实数编码方式。(2) 计算群体中每一个染色体的适应函数值。(3) 判断适应函数中是否存在满足研究问题所需条件的最优解，如果存在或算法停止条件成立则算法停止。计算停止时，如果满足研究问题所需条件则得到相应问题具有最佳染色体的个体，该最佳个体就是所研究问题的最优解。(4) 如果适应函数中没有满足研究问题所需条件的最优解，则根据适应函数值排序，通过非

线性排名选择机制按预定的概率选择一定数量染色体。(5) 将选定的染色体作为父染色体通过杂交的方法产生新的染色体。(6) 对完成杂交的上述种群以预定的概率进行变异;经变异后的种群继续重复第二至第六步。

从遗传算法的求解过程可以看到,与其他算法(如梯度下降法)相比,遗传算法在优化求解时不会落入局部最优且具有很高的搜索效率,因此通过遗传算法可以迅速找到最优解。

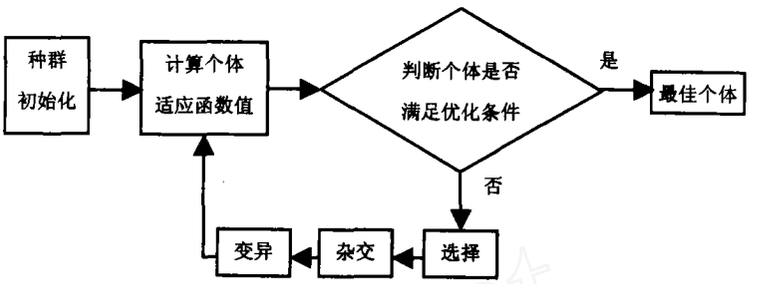

图 1 遗传算法流程图

2. 年龄别生育率与总和生育率间接估计遗传算法优化模型的建立

由于生育孩子数与育龄妇女年龄结构和年龄别生育率之间的相互关系可以用人口预测模型进行表达,而对过去育龄妇女年龄结构和出生人口数则可以通过人口回溯方法得到,因此可以以存活人口倒推得到的出生人口为优化目标,把对年龄别生育率的间接估计问题转化为基于遗传算法的全局优化求解问题,也就是通过遗传算法不断筛选较好的染色体(年龄别生育率),最终找到能够满足人口预测出生人口数与存活倒推出生人口数相吻合的年龄别生育率,具体建模方法是:(1) 根据人口年龄结构间接估计方法估算年龄别育龄妇女数(t_1 时点年龄别育龄妇女人数${}_nPf^{t_1}_{(x)}$)和出生人口数(t_2 时点 0 岁人口数$P'^{t_2}_{(0)}$)。(2) 确定遗传算法中种群初始化染色体${}_nF_{(x)}$(年龄别生育率)的取之范围和染色体发生变异空间解集,如${}_nF_{(x)}\geqslant 0$ 等,以及遗传算法求解的中止条件,如最小误差和最大遗传代数的设定。(3) 根据育龄妇女生育模型建立出生人口、年龄别育龄妇女数和年龄别生育率的函数关系:$P^{t_2}_{(0)}=\frac{L_{(0)}}{2}*\left\{\sum\left[{}_nPf^{t_1}_{(x)}*{}_nF_{(x)}+{}_nPf^{t_1}_{(x)}*{}_nF_{(x+n)}*\frac{{}_nL_{(x+n)}}{{}_nL_{(x)}}\right]\right\}$;式中 $P^{t_2}_{(0)}$是在 t_2 时刻年龄为 0 岁人口数;${}_nPf^{t_1}_{(x)}$是在 t_1 时刻年龄在 x 岁至 x+n 岁的妇女人口数;x 取值范围是 15～49;${}_nF_{(x)}$为年龄在(x,x+n)之间的育龄妇女生育率;${}_nF_{(x+n)}$为年龄在(x+n,x+2n)之间的育龄妇女生育率。(4) 确立遗传算法不同个体的染色体适应函数 $MIN=P^{t_2}_{(0)}-P'^{t_2}_{(0)}$。根据该适应函数对初始种群中染色体进行选择、杂交和变异等遗传进化操作,直到优选出满足最优条件的最佳染色体(育龄妇女年龄别生育率)为止。

(三) 年龄别生育率与总和生育率间接估计实施步骤

由于遗传算法最优解的搜索效率、解的多样性与染色体的初值和约束条件有关,因此对特定出生人口数和特定育龄妇女年龄结构条件下最佳年龄别生育率的搜索空间范围直接影响模型的效率。为了提高遗传算法的搜索效率,本文在进行年龄别生育率和总和生育率间接估计时,采取多次自适应演化计算的方法,具体步骤如下。

第一,根据人口系统的基本特征初步确立解的空间,然后通过多次相同条件下最优求解过程,重复进行求解并得到一系列最优解。在初步搜索过程中,年龄别生育率的取值空间范围可以通过总和生育率的变动范围、Hutterites 人口年龄别生育率和当前年龄别生育率状态来初步确定比较松弛的最优解范围。由于在初步搜索过程中约束条件比较松弛和遗传算法随机模拟解的多样性,因此,可以根据遗传算法优化模型多次反复求解并得到初步约束条件下最优解的置信区间。

第二,在初步求解的基础上,获得人口系统的特性,如与特定年份出生人口、年龄别育龄妇女人口数相对应的总和生育率的取值范围,并根据这些特性进一步缩小约束条件的范围,即根据人口系统的特性缩小解的空间,当然,缩小解的空间是相对于初步搜索过程而言的。求解范围缩小后,解的

空间应该不影响对最优解的搜索和最优解可能的取值范围。通过约束条件的变化，为进一步通过遗传算法在较小的空间内进行最优解的搜索创造条件。

第三，对相应参数进行自适应调整后，使最优解在缩小后的空间内进行搜索，因此可以继续在相同约束条件下重复最优求解过程，多次求解并得到相应的最优解。

第四，运用统计方法，得到年龄别生育率和总和生育率最优解的均值和置信区间。需要说明的是，之所以采用大量最优解的均值，是因为上述对年龄别生育率和总和生育率的约束条件比较松弛，因此容易发生由于受随机因素的影响，极个别年龄组解的不稳定问题。也就是说，如果只进行一次求解得到的结果经常发生绝大多数年龄别生育率与目标高度吻合，而极少数年龄组年龄别生育率在约束条件内与实际目标明显不符并违反一般生育规律的震荡的问题，从而产生奇异点问题，如40岁年龄别生育率大于25岁年龄别生育率等。

二、年龄别生育率与总和生育率间接估计方法应用实例

运用人口普查数据，可以对普查以前育龄妇女年龄别生育率与总和生育率进行间接估计，具体间接估计的年份可根据需要确定。本文以1990年人口普查数据为例，通过对1982年全国年龄别生育率与总和生育率间接估计来展示上述间接估计方法数据需求、基本参数设定和实施过程。

首先，进行数据准备和基本参数初始化。人口数据准备包括1990年男性、女性年龄别人口数，男性、女性年龄别死亡率和1990年育龄妇女年龄别生育率。遗传算法的基本参数设定包括种群规模为100，选择杂交概率为0.2，变异概率为0.5，最大遗传代数为400，最优解的数量为500。

其次，进行初步间接估计，确定解的空间。在初步间接估计年龄别生育率和总和生育率时，将解的空间范围假定在1990年年龄别生育率的4倍和1/4之间，遗传算法求解发生变异的空间在1990年年龄别生育率的8倍和1/8之间。可以确信1982年年龄别生育率和总和生育率的解集一定在上述空间之内。经过初步求解，得到1982年总和生育率的99%置信区间为[2.96±0.16]。

第三，根据初步估计，进一步收缩解的空间。将解的空间范围假定为1990年年龄别生育率($ASFR(x)_{1990}$)的($1.5 * TFR_{上限}/TFR_{1990}$)倍和($TFR_{1990}/(1.5 * TFR_{上限})$)之间，遗传算法求解发生变异的空间在1990年年龄别生育率的($3 * TFR_{上限}/TFR_{1990}$)倍和($TFR_{1990}/(3 * TFR_{上限})$)之间。具体解空间和变异空间见图2和图3，可以确信1982年年龄别生育率的取值将在图3所描述的解的变异空间之内。基于上述参数和解的空间，得到一系列最优解，最优解的均值作为对年龄别生育率和总和生育率的间接估计值。对1982年年龄别生育率的间接估计结果见图4。

通过对年龄别生育率和总和生育率的间接估计，不但可以比较确切地反映各年度生育模式和生育水平的基本状况，而且可以反映其动态变化趋势。例如对1986年年龄别生育率的估计(见图

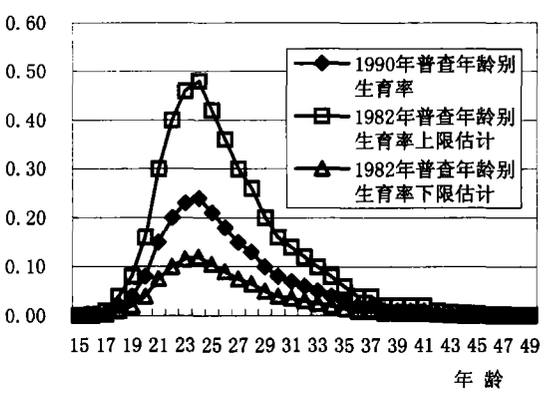

图2 1982年年龄别生育率最优解空间范围估计

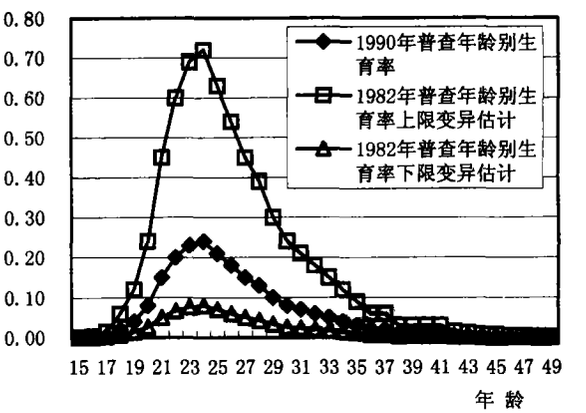

图3 1982年年龄别生育率最优解变异空间范围估计

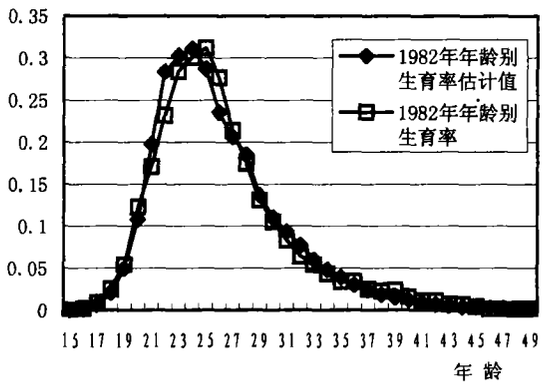

图 4　1982 年年龄别生育率间接估计图

注：根据 1990 年人口普查数据推算，1982 年年龄别生育率来源于姚新武编：《中国生育数据集》，中国人口出版社，1995 年。

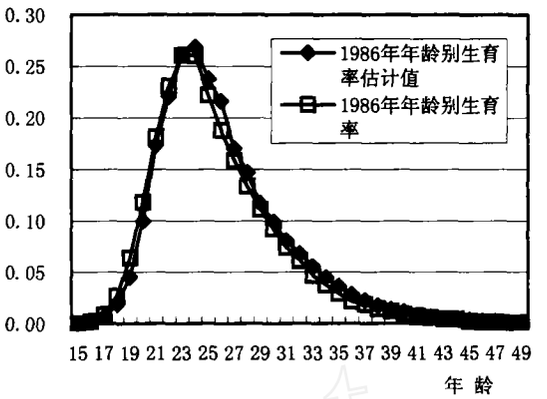

图 5　1986 年年龄别生育率间接估计图

注：根据 1990 年人口普查数据推算，1986 年年龄别生育率来源于姚新武编：《中国生育数据集》，中国人口出版社，1995 年。

表　1982～1989 年总和生育率估计值

估计值	1982 年	1983 年	1984 年	1985 年	1986 年	1987 年	1988 年	1989 年
总和生育率	2.89	2.64	2.44	2.45	2.49	2.68	2.55	2.35
估计值上限*	2.98	2.73	2.55	2.60	2.67	2.85	2.72	2.47
估计值下限*	2.80	2.55	2.34	2.29	2.31	2.50	2.39	2.24

注：根据 1990 年人口普查数据推算。* 99%置信水平。

5)，就可以充分反映当时生育模式和生育水平的状况和变化。同样，根据 1990 年人口普查数据和本文算法可以获得 1990 年以前各个年份总和生育率的变动情况和统计置信区间(见表)。

从以上实例我们可以看到，育龄妇女年龄别生育率与总和生育率间接估计方法可以比较确切地反映生育模式和生育水平的可变动趋势。但对于完全对应各年龄组年龄别生育率的准确估计还需要进一步努力，如对 1982 年年龄别生育率的间接估计。之所以产生误差，一方面由于原始数据的质量，另一方面说明对年龄别生育率解的空间需要通过人口学内在规律进一步约束。尤其是生育模式的内在规律无疑会使间接估计的准确程度进一步提高。此外，时间期限与原始数据初始年度的差距越大，人口年龄结构间接估计时的累计误差越大，因此，对各年度人口存活状况的估计误差也成为年龄别生育率间接估计误差的来源之一。可见，同时引入其他优化与间接估计方法也将成为提高对育龄妇女年龄别生育率与总和生育率间接估计精度的重要努力方向。

参考文献：

1. 郭志刚：《时期生育水平指标的回顾与分析》，《人口与经济》，2000 年第 1 期。
2. 王广州：《人口年龄结构间接估计方法与应用研究》，《中国人口科学》，2001 年第 5 期。
3. 国务院人口普查办公室、国家统计局人口统计司编：《中国 1982 年人口普查资料》、《中国 1990 年人口普查资料》，中国统计出版社，1985 年、1993 年。
4. Bongaarts, J. and G. Feeney (1998), On the Quantum and Tempo of Fertility, *Population and Development Review* 24(2), 271－291.
5. Samuel H. Preston, Patrick Heuveline and Michel Guillot(2001), Demography－Measuring and Modeling Population Processes, Blackwell Publishers Ltd.
6. David P. Smith(1992), Formal Demography, Plenum Press.

(责任编辑：朱　萍)