利用生育模型及队列与时期指标的转换对生育率预测的方法学研究

谢韦克　王绍贤

【摘要】　本文讨论了有全部生育率数据、只有部分生育率数据以及没有生育率数据3种不同条件下构造生育模型的方法及意义。首次指出构造Brass相关生育模型的非最小二乘法（曾毅方法）优于最小二乘法，并对其原因进行了分析。

【作者】　谢韦克　北京医科大学卫生统计与医学人口教研室，副教授；王绍贤　北京医科大学卫生统计与医学人口教研室，教授。

生育模型的理论在生育分析及生育规划中有重要作用。如何在不同的条件下构造生育模型，使之能更准确地用于政策比较与生育预测，一直是人口分析技术中的重要课题，本文利用前人的研究结果，并借助队列—时期生育率矩阵，着重就方法学方面的问题作进一步的探讨。

1. 不同条件下构造生育模型的方法及意义

一般，我们是在3种不同的条件下构造生育模型：（1）有全部年龄组的生育率数据；（2）有部分年龄组的生育率数据；（3）没有生育率数据。不同条件下构造生育模型的方法及作用各不相同，我们分别进行讨论。

1.1 有全部生育率数据的条件下构造生育模型

如果有全部生育率数据，一般用最小二乘法确定模型中的参数。与其它方法相比，最小二乘法可以达到最好的拟合效果。由于是对现有的全部生育率数据进行曲线拟合，这种建立生育模型的方法与生育预测并没有直接的关系。它的目的，主要是从拟合效果的角度检验所使用的函数形式能否较好地反映生育率随年龄变化的规律。到目前为止，国内外人口学家已提出不少反映生育模式的函数形式。当然，不同地区、不同民族及不同时期的生育模式是多种多样的，不可能用一种函数形式来概括。我们可以利用最小二乘法研究每种函数的适用范围，对现有的函数进行改进。同时，也可以对诸多的函数形式进行比较研究，寻找拟合效果最好、适用范围最广，特别是能适用于中国当今计划生育政策下生育模式的函数形式。然而，在第一种情况下用最小二乘法构造生育模型似乎存在着一种矛盾。一方面，只有当某种函数对现有的生育率数据能有很好的拟合效果，才可能利用它进行生育预测与生育分析；另一方面，某一函数对现有的生育率数据拟合得再好，也不能代替未来的生育模式。孙以苹(1986)曾经指出：如果我们根据1972年统计数据构造生育模式，无论拟合情况怎样好，也很难满意地反映出1973年的情况。对此，他提出了力求用较少而又容易得到的资料去构造生育模式的方法。这也就是第二种情况，利用部分年龄组的生育率数据构造生育模型。

1.2 只有部分生育率数据的条件下构造生育模型

利用部分年龄组的生育率数据构造生育模型，需要解决3个问题：方法学及方法的效果，

* 本文系国家计划生育委员会八五期间第二批人口与计划生育研究课题。

部分年龄组的生育率数据能少到什么程度，能在多大程度上解决总和生育率的近期预测。

关于第一个问题，如果存在构造生育模型的非最小二乘法，也即模型中的参数与具有人口学意义的指标之间存在解析表达式，则问题就转化为能否利用部分生育率数据确定具有人口学意义的指标。一般，对于所直接拟合的曲线来说，最小二乘法的拟合效果要优于非最小二乘法，有时两种方法的拟合效果可以相差很多。比如，宋健等（1985）利用X^2-分布拟合江苏省如东县1973年生育率数据。X^2-分布只有1个参数n，如果用最小二乘法，参数n＝6。如果用非最小二乘法，根据生育率数据可知$a_o=16$，$a_{max}=23$，再根据公式$a_{max}=a_o+n-2$可解得n＝9。两种方法下对参数估计的相对误差高达50%。乔晓春（1991）也指出，利用峰值生育年龄与初始生育年龄的差构造X^2-分布的生育模型会带来严重的错误。因此，当某种函数形式用最小二乘法去拟合生育率数据能有较好的拟合效果时，还必须研究其非最小二乘法的拟合效果。笔者曾经用3种方法：最小二乘法、非最小二乘法，以及在部分生育率数据条件下的非最小二乘法，对1989年生育率数据进行对数正态分布的曲线拟合，以考察不同条件与不同方法下的拟合效果（谢韦克，1993）。由于Brass相关生育模型在中国人口学界有比较广泛的应用，因此本文就Brass相关生育模型的非最小二乘法进行探讨。

关于构造Brass相关生育模型的非最小二乘法，曾毅（1991）提出了利用生育年龄的中位数及第三与第一个四分位数之差求取α和β的方法。假定1989年生育模式服从Brass相关生育模型，张二力等（1994）已经建立了新的一套标准生育模式及其转换值，并利用最小二乘法估计出模型中的参数。本文利用同一个标准生育模式，用曾毅教授提出的方法估计模型中的参数α和β。两种方法的结果及拟合效果见表1。表1中$\sum(Y_i-\widehat{Y}_i)^2$为误差平方和，$Y_i$为年龄别生育率的观察值，$\widehat{Y}_i$为拟合值。

表1　两种方法构造Brass相关生育模型的结果

胎次	最小二乘法			非最小二乘法		
	α	β	$\Sigma(Yi-\hat{Y}i)^2$	α	β	$\Sigma(Yi-\hat{Y}i)^2$
总和	0.123110	0.953487	0.00025	0.138572	0.953650	0.00024
一胎	0.100916	0.916024	0.00210	0.194192	1.024159	0.00042
二胎	−0.021259	0.806937	0.00167	0.078826	0.876899	0.00071
三胎+	0.088139	0.904551	0.00062	0.071073	0.992832	0.00019

资料来源：中国1990年人口普查资料，第3册。

从表1可以看出，用Brass相关生育模型拟合1989年生育率，非最小二乘法的拟合效果明显优于最小二乘法，其误差平方和均为万分位数。进一步的分析表明，非最小二乘法拟合误差的减少主要表现在峰值生育率的附近，参见图1、2。对Γ-分布或对数正态分布的生育模型来说，非最小二乘法的拟合效果不可能优于最小二乘法。而Brass相关生育模型的非最小二乘法却可以优于最小二乘法，这是由构造Brass相关生育模型的特殊性造成的，限于篇幅，本文只进行简单的分析。

构造Brass相关生育模型并不直接拟合生育率随年龄变化的曲线，而是在标准累计生育率的转换值Ys(x)与所研究的累计生育率的转换值Y(x)之间拟合一条直线Y(x)＝α＋βYs(x)。对于这条直线方程来说，非最小二乘法的拟合效果不会优于最小二乘法。利用1989年生育率数据计算的结果也是如此。但是，直线方程的拟合效果与转换成生育率曲线的拟合效果并不

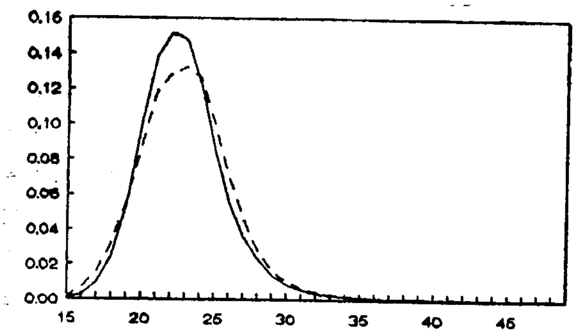

图1 1989年1孩生育模式的观察值（实线）与模型值（虚线），最小二乘法拟合结果

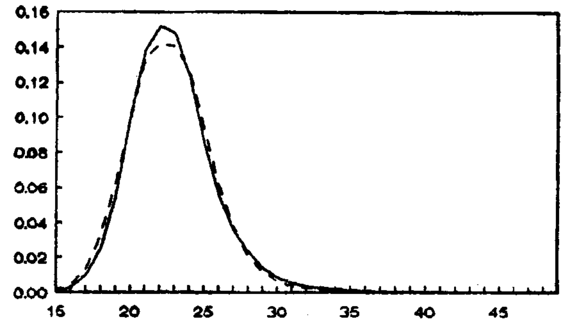

图2 1989年1孩生育模式的观察值（实线）与模型值（虚线），非最小二乘法拟合结果

等价。若令Y(x)为各年龄组累计生育率两次对数转换后的转换值，a_m为峰值生育年龄，则Y(x)的绝对值从15岁开始随年龄的增加而减少，一般在a_m+1岁达到最小，以后又逐渐增大，在49岁达到最大（张二力等，1994）。也就是说，生育率越大（小）的年龄组，其Y(x)的绝对值越小（大）；反之亦然。因此，直线方程中不同年龄组Y(x)的拟合误差并不具有同等的意义。高生育率年龄组Y(x)的误差的减少（或增加）与低生育率年龄组Y(x)相同误差的减少（或增加）相比，能造成生育率拟合误差更多的减少（或增加），所以，减少峰值生育年龄附近Y(x)的拟合误差就具有重要的意义。

Y(x)的分布不是对称的，其平均值明显偏向高年龄组。由于最小二乘法确定直线方程要考虑所有的数据，而且该直线方程一定会通过点（$\bar{y}(x)$，$\bar{y}_s(x)$）。因此，直线方程一般会在y(x)的平均值$\bar{y}$(x)的周围有最好的拟合效果。在利用生育年龄的中位数及第一三个四分位数确定直线方程时，由于这3个生育年龄的百分位数都在峰值生育年龄附近，因此，非最小二乘法确定的直线方程一般会对峰值生育年龄附近的Y(x)有最好的拟合效果。如前所述，这会大大增加生育率曲线的拟合优度，这也就是非最小二乘法构造Brass相关生育模型可以优于最小二乘法的原因。

构造Barss相关生育模型的非最小二乘法可以优于最小二乘法，应该说是其一大优点，但是否在任何情况下都成立，与所选取的标准生育模式是否有关？有的研究者还提出了其它形式的非最小二乘法（张二力、陈建利，1994)，其拟合效果是否也优于最小二乘法？还有，能否对现行的最小二乘法实行某些改进，以提高直线方程对峰值生育年龄周围的Y(x)的拟合精度？笔者认为，这些都是值得进一步深入讨论的问题。

关于第二个问题，即构造生育模型对生育率数据的最低要求，在生育模式服从Γ-分布的假定下，孙以苹提出了利用峰值生育年龄以及峰值生育率估计参数的方法。该方法要求知道生育水平也即总和生育率（孙以苹，1986）。笔者曾经指出，无论是Γ-分布还是对数正态分布的生育模型，只要有生育率数据的年龄组包含峰值生育年龄与平均生育年龄，就可以同时估计模型中的参数以及生育水平（谢韦克，1993）。总之，两种方法都要求至少知道峰值生育年龄。用非最小二乘法构造Brass相关生育模型，需要计算生育年龄的中位数及第一与第三个四分位数。与平均生育年龄、峰值生育年龄的计算与确定不同，如果缺少两头的生育率数据，平均生育年龄与峰值生育年龄仍然可以计算与确定。而确定生育年龄的百分位数必须计算累计生育率与总和（或终身）生育率之比。因此，计算上述3个生育年龄的百分位数

必须知道总和生育率的值以及至少75%的年龄别生育率。这与百分之百生育率的要求条件已相差不远。应该说，非最小二乘法构造Brass相关生育模型对生育率数据的要求是最严格的。

关于第三个问题，即这种利用部分生育率数据构造生育模型的方法，能在多大程度上解决总和生育率的近期预测。如果某一队列妇女已完成她们的峰值生育，则我们可以估计该队列妇女峰值生育年龄以后的年龄别生育率。根据队列指标与时期指标的转换，我们可以得到预测年的一个年龄别生育率。如果某一队列妇女还未达到峰值生育年龄，我们不能利用她们过去的生育率数据构造出生育模型以估计她们以后的年龄别生育率，也就不能得到预测年总和生育率中与该队列妇女相对应的年龄别生育率。因此，这种利用部分生育率数据构造生育模型的方法，只能部分地解决总和生育率的近期预测。关于这个问题，我们在后面借助队列一时期生育率矩阵进行更详细的讨论。

顺便说一句，这种利用部分生育率数据构造生育模型的方法，在总和生育率的间接估计中有一定的作用。比如，如果某地区总和生育率的调查，由于瞒报或误报使低年龄组妇女的生育资料不易得到或质量较差，则我们可以利用峰值生育年龄附近高质量的生育率数据，用构造生育模型的方法对低年龄组妇女的生育率进行间接估计。需要指出的是，这与生育率的预测是完全不同的两个概念。

1.3 没有生育率数据的条件下构造生育模型

在没有生育率数据的条件下构造生育模型，一般用于某些理论分析或政策仿真，也可用于人口的长期规划或预测。当现有的生育率数据不足以构造生育模型时，比如要估计还未达到峰值生育的队列妇女以后的生育模式，也属于这种情况。第三种情况下同样用非最小二乘法，即利用人口学指标与模型中参数的关系构造生育模型。与第二种情况不同的是，这里人口学指标的确定不是根据生育率数据。比如，如果要讨论逐步提高平均生育年龄对中国人口发展的影响，这里平均生育年龄的确定或者根据其本身的变动趋势，或者是某种计划生育政策要求下的理论假设。还有，第三种情况下构造生育模型只能确定生育模式而不能确定生育水平，如果要进行年龄别生育率的预测，还需要更多的假设条件，其预测的价值自然就较低。

在第二种情况下，我们从拟合效果的角度，以Brass相关生育模型为例讨论了非最小二乘法。在第三种情况下，我们根据生育模式的变动与人口学指标变动间的关系讨论非最小二乘法。

如果构造Γ-分布或对数正态分布的生育模型，需要知道平均生育年龄、峰值生育年龄以及生育年龄的标准差这3个指标中的任意2个；如果构造Brass相关生育模型，需要知道生育年龄的中位数以及第一三个四分位数3个指标。当生育模式发生变化时，人口学指标也发生相应的变化，因此我们才有可能利用人口学指标进行生育预测。关于峰值生育年龄，孙以苹、乔晓春都指出，自60年代至80年代初，总和生育率的峰值生育年龄基本不变，但生育模式却发生了巨大变化。因此，利用峰值生育年龄进行生育预测也就有一定的困难（孙以苹，1986；乔小春，1991）。尽管上述情况在80年代后期发生了变化，但峰值生育年龄对于生育模式的变化确实不十分敏感。但应该指出：首先，如果我们考察分胎次的生育，则上述情况并不成立；其次，对于总和生育率或相同胎次别的生育率来说，平均生育年龄与生育年龄的中位数也不可能在一年的时间发生一个整岁的变化（参见表2）。又比如，中国城市1980年和1988年35岁同批妇女的一胎平均生育年龄分别为22.15和23.89岁。8年时间提高了1.74岁，平均每年增长0.22岁，最大增长幅度为0.30岁（谢韦克，1994）。

由于平均生育年龄与生育年龄中位数的计算利用了全部生育率数据，因此小于一个整岁

表2 不同时期总和生育率的3个人口学指标的变化

年份	峰值生育年龄	平均生育年龄	生育年龄的中位数
1964	25	29.68	28.00
1974	25	29.18	27.02
1980	25	27.41	25.39
1987	24	26.19	24.17
1989	23	26.11	25.08

资料来源：全国生育节育抽样调查全国数据卷；中国1990年人口普查资料，第3册。

的变化也可以计算出来。而峰值生育年龄的确定只利用一个年龄组的生育率，因此都是整数岁，生育模式的变化也就难以在峰值生育年龄上表现出来。如果生育模式能较好地服从对数正态分布，则利用对数正态分布可计算出非整数值的峰值生育年龄。我们可以利用生育模型的理论对峰值生育年龄进行比较精确的估计，以研究峰值生育年龄的变动趋势。对于Γ-分布与对数正态分布的生育模型，峰值生育年龄必须和平均生育年龄结合起来才能确定模型中的参数，因此可以利用平均生育年龄的变动以弥补峰值生育年龄“不变”的不足。

中国生育模式的变化与人口学指标变化间的关系还表明，当平均生育年龄提高时，不论其提高幅度的大小，不论起始生育年龄是否发生变化，也不论是时期指标还是队列指标，生育模式的变化都表现为生育率曲线形状的变化，而并不表现为生育率曲线的平移。曾毅(1993)提出的平均生育年龄提高时计算年龄别生育率的插值公式，实际上是初始生育年龄提高而生育模式不变条件下的插值公式，正如其所指出的，这会产生一些误差。比较准确的方法是利用平均生育年龄与其它人口学指标构造出生育模型，以讨论逐步提高平均生育年龄对中国人口发展的影响。

2. 利用队列—时期生育率矩阵进行生育预测及生育分析

为了讨论的方便，引进队列一时期生育率矩阵，简称CP矩阵。

$$CP=\begin{pmatrix} a_{0\ 15} & a_{-1\ 15} & a_{-2\ 15} & \cdots & a_{-34\ 15} \\ & a_{0\ 16} & a_{-1\ 16} & \cdots & a_{-33\ 16} \\ & & a_{0\ 17} & \cdots & a_{-32\ 17} \\ & & & \ddots & \vdots \\ & & & & a_{-1\ 48} \\ & & & & a_{0\ 49} \end{pmatrix}$$

CP矩阵是一个上三角形矩阵。第1列到第35列分别是调查年15～49岁队列妇女的年龄别生育率。矩阵中a_{0x}表示调查年x岁队列妇女x岁的生育率，a_{-ix}表示调查年前i年x岁队列妇女x岁的生育率。比如，$a_{0\ 17}$表示调查年17岁队列妇女17岁的生育率，$a_{-1\ 16}$表示调查年前1年16岁队列妇女16岁的生育率，也是调查年17岁队列妇女16岁时的年龄别生育率。矩阵中主对角线元素正好是调查年的总和生育率。显然，CP矩阵是生育调查年所能得到的关于15～49岁队列妇女生育率的最详尽的资料。假定调查年为t年。到了t+1年，CP矩阵中第35列退出矩阵，t+1年15岁妇女进入矩阵。其它队列妇女增加一个年龄别生育率并前移1列。t+1年的CP矩阵只有主对角线元素为新增加的元素，其余元素均为t年CP矩阵中的元素。显然，t+1年总和生育率（不包括15岁）是t年CP矩阵34个队列妇女的生育率发展变化的结果。如果利用t年总和生育率推测t+1年总和生育率，实际上只利用了t年的CP矩阵中主对角线元素。如果利用CP矩阵所有的元素进行生育预测与生育分析，就可以最充分地利用生育率数据，更好的反

映生育率的变动趋势，也可以清楚地反映队列与时期指标间的转换关系。为了究研的需要，还可以构造胎次别CP矩阵，分性别CP矩阵、初婚率CP矩阵等等。表3给出了1987年24～29岁队列妇女的年龄别生育率。从中可以看出队列妇女生育率的变动趋势。比如，峰值生育率随队列妇女出生年代的后移而减少，但从25岁队列妇女又开始呈回升趋势。

表3 1987年24～29岁6个队列妇女的年龄别生育率

24	25	26	27	28	29
0.0002	0.0002	0.0007	0.0004	0.0010	0.0006
0.0016	0.0017	0.0019	0.0018	0.0029	0.0029
0.0050	0.0067	0.0053	0.0049	0.0066	0.0077
0.0194	0.0186	0.0175	0.0121	0.0158	0.0208
0.0545	0.0481	0.0370	0.0337	0.0333	0.0338
0.1054	0.1232	0.0824	0.0622	0.0696	0.0568
0.1679	0.1943	0.1709	0.1232	0.0969	0.1036
0.2039	0.2491	0.2077	0.2319	0.1856	0.1320
0.2561	0.2691	0.2498	0.2399	0.2841	0.2270
0.2695	0.2936	0.2444	0.2566	0.2740	0.3012
	0.2626	0.2283	0.2117	0.2429	0.2447
		0.1976	0.1901	0.1858	0.1947
			0.1606	0.1590	0.1468
				0.1438	0.1314
					0.1136

资料来源：全国生育节育抽样调查全国数据卷。

前面已经提到，对于已经完成峰值生育的队列妇女，可以利用伽玛分布或对数正态分布，估计其峰值生育年龄以后的年龄别生育率。假定CP矩阵中a_{024}是该队列妇女的峰值生育率，再假定CP矩阵中凡大于24岁的队列妇女均已完成峰值生育，则我们可以估计CP矩阵中所有大于24岁的队列妇女在调查年（t年）以后的年龄别生育率，即利用生育模型的理论“帮助”她们完成一生的生育。根据队列指标与时期指标间的转换关系，我们可以得到t＋1年的总和生育率中25～49岁的年龄别生育率、t＋2年的总和生育率中的26～49岁的年龄别生育率，余类推。预测年越远，CP矩阵中可以利用的数据就越少，所得到的预测年的年龄别生育率也越少。

构造那些已完成峰值生育的队列妇女的生育模型，就必须计算她们的平均生育年龄与峰值生育年龄。由此可得到反映生育模式的人口学指标的变动趋势，因此，对于那些未达到峰值生育的队列妇女，我们可以利用人口学指标的变动趋势，用构造生育模型的方法估计她们未来的生育模式，在某一生育水平的假定下计算未来的年龄别生育率。

我们可以把CP矩阵分为两部分，一部分是已完成峰值生育的队列妇女，我们可以根据她们以往的生育率估计她们未来的生育率。由于她们已完成峰值生育，她们的生育在以后可以变动的范围也较小，预测的可信度也就较高。另一部分是未达到峰值生育的妇女，我们不能利用她们过去的生育率构造出生育模型，只能用其它方法对生育率进行预测。由于需要更多的假设条件，预测的可信度也就较低。由于她们还未达到峰值生育，她们的生育在以后可以变动的范围就较大。我们可以根据目前的生育模式及生育水平的变动趋势估计她们以后的生育，也可以在人口规划的条件下“规定”她们的生育模式及生育水平，计算不同生育模式及不同生育水平下未来的出生数，以进行政策仿真与比较。

与直接构造预测年的生育模型相比，这种利用队列与时期指标的转换对生育预测的方法比较复杂。但在某些情况下，比如，如果要讨论逐步提高平均生育年龄对中国人口发展的影响，就只能用这种方法。对于时期指标来说，两个具有不同平均生育年龄的总和生育率可以有相同的生育水平。而对于队列指标来说，具有相同生育水平的两个队列妇女，如果她们的平均生育年龄不同，她们所对应的时期指标肯定不一样。

3. 实例分析

1988年2‰生育节育调查提供了比较丰富的队列妇女的生育资料。用笔者曾提出（1994）的方法，利用1987年25～29岁5个队列妇女的生育率，预测1989年时期指标总和生育率中27～31岁年龄别生育率，并与1990年全国第四次人口普查得到的生育率比较（见表4）。

表4 1989年部分年龄别生育率的预测值与观察值的比较

年龄	胎次	生育率		
		预测值	观察值	相对误差
27	总和	0.1933	0.1467	0.3176
	一胎	0.0387	0.0358	0.0810
28	总和	0.1316	0.1140	0.1544
	一胎	0.0204	0.0240	0.1500
29	总和	0.1020	0.0966	0.0559
	一胎	0.0137	0.0145	0.0551
	二胎	0.0649	0.0404	0.6064
30	总和	0.0736	0.0813	0.0947
	一胎	0.0085	0.0085	0.0000
	二胎	0.0393	0.0337	0.1661
31	总和	0.0645	0.0673	0.0416
	一胎	0.0119	0.0054	1.2037
	二胎	0.0305	0.0267	0.1423

资料来源：同表2。

从表4可以看出，13个指标中有3个指标预测值的误差很大，相对误差为31.76%～120.37%；4个指标预测值的误差较大，为15%左右；6个指标预测值的误差较小，10%以下。即有将近一半的预测值效果不理想。笔者认为，造成这种误差的主要原因是资料问题。如果以1990年人口普查10%抽样资料中的生育率做为假想队列妇女的生育率，则利用部分生育率数据估计峰值生育以后的年龄别生育率及总的生育水平可以得到很好的结果。而利用2‰抽样资料得到的生育率的预测值与普查值比较，抽样误差是难以避免的。这也说明了利用CP矩阵对总和生育率进行近期预测离实用还比较遥远，因为我们很难迅速得到足够数量的队列妇女的生育率资料。

笔者认为，利用CP矩阵主要在方法学上有较大的意义，我们可以利用CP矩阵深入分析队列与时期指标间的转换关系。比如，初育间隔、生育间隔、生育胎次递进比等队列指标的变动如何影响总和生育率的变动。又比如，当缺少队列资料时，有些队列指标的计算只能借助时期资料，这会造成什么样的误差？这些问题，如果借助CP矩阵都可以得到比较充分的讨论。对此，笔者准备另文做专门的讨论。

参考文献

1 孙以苹.非固定生育模式构造方法的探索.人口与经济，1986.3

2 宋健，于景元.人口控制论.科学出版社，1985

3 乔晓春.对Γ-分布生育模式的讨论.人口研究，1991.6

4 谢韦克.对概率分布生育模型的探讨.中国人口科学，1993.6

5 曾毅等.中国女性婚后离家模型.中国人口科学，1991.1

6 张二力，陈建利.改进的Brass相关生育模型.中国人口科学，1994.6

7 谢韦克等.试论晚育对延缓我国人口增长的作用.南方人口，1994.2

8 曾毅.人口分析方法与应用.北京大学出版社，1993

（本文责任编辑：徐　莉）