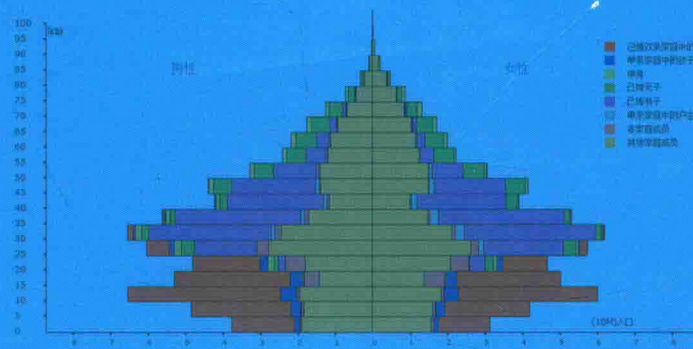
多状态人口预测模型：方法、原理及应用

国家「十二五」科技支撑计划项目
「人口与发展数学模型与综合决策支持系统研究系列」专著之三

国家出版基金项目
NATIONAL PUBLICATION FOUNDATION

Multi-state Model of Population Prospects: Method, Principle and Application

陈佳鹏　等著



中国人口出版社
China Population Publishing House
全国百佳出版单位

责任编辑　周炳然
　　商成果
美术编辑　刘海刚
责任印制　林　鑫
ISBN 978-7-5101-3331-2
定 价 60.00元

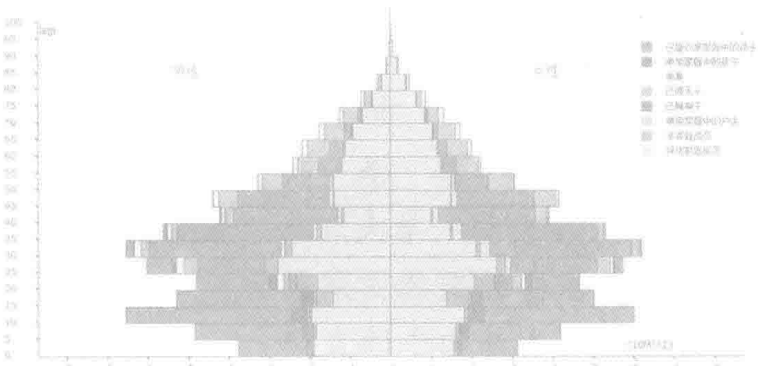
国家出版基金项目
NATIONAL PUBLICATION FOUNDATION

多状态人口预测模型：方法、原理及应用

国家『十二五』科技支撑计划项目
『人口与发展数学模型与综合决策支持系统研究系列』专著之三

Multi-state Model of Population Prospects: Method, Principle and Application

陈佳鹏　等著



中国人口出版社
China Population Publishing House
全国百佳出版单位

图书在版编目(CIP)数据

多状态人口预测模型:方法、原理及应用/陈佳鹏等著.
—北京:中国人口出版社,2016.12
(国家"十二五"科技支撑计划项目"人口与发展数学模型与综合决策支持系统研究系列"专著)
ISBN 978-7-5101-3331-2

Ⅰ.①多… Ⅱ.①陈… Ⅲ.①人口预测-研究-中国
Ⅳ.①C924.23

中国版本图书馆 CIP 数据核字(2015)第 092801 号

多状态人口预测模型:方法、原理及应用
陈佳鹏 等著

出 版 发 行 中国人口出版社
印 刷 北京中印联印务有限公司
开 本 787 毫米×1092 毫米 1/16
印 张 22
字 数 400 千字
版 次 2016 年 12 月第 1 版
印 次 2016 年 12 月第 1 次印刷
书 号 ISBN 978-7-5101-3331-2
定 价 60.00 元

出 版 人 邱 立
网 址 www.rkcbs.net
电子信箱 rkcbs@126.com
总编室电话 (010)83519392
发行部电话 (010)83530809
传 真 (010)83519401
地 址 北京市西城区广安门南街 80 号中加大厦
邮 编 100054

国家“十二五”科技支撑计划项目
“人口与发展数学模型与综合决策支持系统”

项目负责人 蒋正华

课题负责人 姜卫平

项目执行委员会成员

姜卫平 沈丽文 刘晓丽 刘鸿雁 武家华 张许颖
陈佳鹏 庄亚儿 王 宁 李沛霖(课题秘书)

“多状态人口分析预测关键技术与模型研究”子课题

子课题负责人 陈佳鹏

研究团队成员 刘中一 龚双燕 陈 卓 李成福
王 勇 刘冬梅 张 莉 蔚志新
史 卓 麻 薇 王 哲 张翠玲
陈 恩

技术团队成员 丁志恒 高 宁 周立权 杨 鑫
刘 彤 霍 伟

专家支持团队 北京大学 郑晓瑛 乔晓春 陈 功 张 蕾
国家发改委社会发展研究所 杨宜勇
教育部中国教育科学研究院 马晓强
中国老龄科研中心 王海涛
国家卫生计生委卫生统计信息中心 王才有 徐 玲

总　序

人是社会发展的动力，也是社会发展的主体。对于人口与经济社会发展内在规律的探究是一个古老课题。古代人类生产力低下，资源供给丰富，利用方式粗放，生活环境艰苦，死亡率极高，需要高生育率以维持种群的存在及发展。进入工业化时期，生产力空前提高，资源匮乏、生态失调、环境破坏等问题不断出现，人们逐渐认识到人口与经济、社会、资源、环境等协调发展的重要性。然而，人类社会是一个复杂的生态系统，个人、家庭、国家等不同主体在人口问题上有不同的利益关系和发展要求，要实现一个为大家普遍接受的均衡状态绝非易事，甚至是世界性难题。解决这个难题需要构建人口发展综合决策系统，把人口和相关因素放在一起进行统筹研判，实现人口发展决策的科学化和智能化。这无疑是一项意义重大而难度超乎想象的工作。

为此，长期以来，国内外人口学和其他众多领域的学者孜孜以求，付出了大量心血。由中国人口与发展研究中心承担的“十二五”国家科技支撑计划项目“人口与发展数学模型与综合决策支持系统”，正是一项具有代表性的研究工作。该项目是首个由国家科技支撑计划资助的人文社科项目，目标是创建具有世界先进水平的人口与发展综合决策支持系统。经过二百多人的研发团队两年多的努力，最终建成以人口模型为核心，集成经济、社会、资源、环境、能源等领域先进成熟模型的人口与发展综合决策支持系统。现在呈现在读者面前的是该项目的部分成果，内容涵盖模型生命表拓展、多状态人口分析预测、多区域人口迁移流动预测、人口政策宏观预测和微观仿真、人口与发展综合决策系统等关键技术。这些成果基于中国人口发展领域的实践，吸收借鉴了国内外相关研究领域的新理论、新方法和新技术，具有国际前沿水平。

我们正身处一个人口态势急剧变化和日趋复杂的世界。人口增长仍

在持续,人口流动更加频繁,社会生活日益多样化,人文内涵继续进化,城市化率快速提高,人口老龄化的挑战前所未有。这些问题,既是中国的也是世界的,加强人口综合科学决策是世界各国尤其是发展中国家的共同需求。希望我们的研究成果能够为此发挥作用。当然,目前的成果仍然是初步的,尚需在实践中加以改进和完善。期待广大科研工作者为此共同努力。

是以为序。

蒋心华

2015 年 5 月

前　言

多状态人口预测系统是“人口与发展数学模型与综合决策支持系统”的主要研究任务之一，目标是开发多状态人口分析预测技术，对家庭、教育、就业、收入、健康、养老等人口状态进行精细分析和综合预测。

多状态人口预测系统研发团队共18人，他们是陈佳鹏、陈卓、刘中一、龚双燕、李成福、王勇、刘冬梅、张莉、蔚志新、张翠玲、麻薇、史卓、王哲、陈恩、丁志恒、周立权、刘彤、高宁，其中5位副研究员、8名博士、3名工程师。具体由5名骨干研究人员牵头分别进行相关多状态模型的研究工作，其他成员配合。

1. 教育多状态模型：陈　卓、刘冬梅
2. 婚姻多状态模型：刘中一、张　莉
3. 健康多状态模型：李成福
4. 家庭多状态模型：龚双燕
5. 就业多状态模型：王　勇

多状态人口预测系统分解成教育多状态、健康多状态、婚姻多状态、家庭多状态、就业多状态等模块，研究框架见图1，经过近3年的研究，目前已完成软件开发和应用，形成婚姻、家庭、健康、教育、就业等多状态人口预测分析的技术标准、实现方案和软件系统，形成了涵盖人生五个重要阶段的多状态人口预测模型体系，见图2。

这是国内首次将多状态生命表与人口预测技术融合，构建了多状态人口预测模型体系，可对婚姻、家庭、健康、教育、就业等多状态人口事件进行中长期预测和精细分析。

由于多状态人口预测是较为新颖的技术方法，受研发团队的知识结构等因素限制，当前的软件系统难免存在各种不足，希望能在今后的应用中不断完善，也欢迎各位读者批评指正。

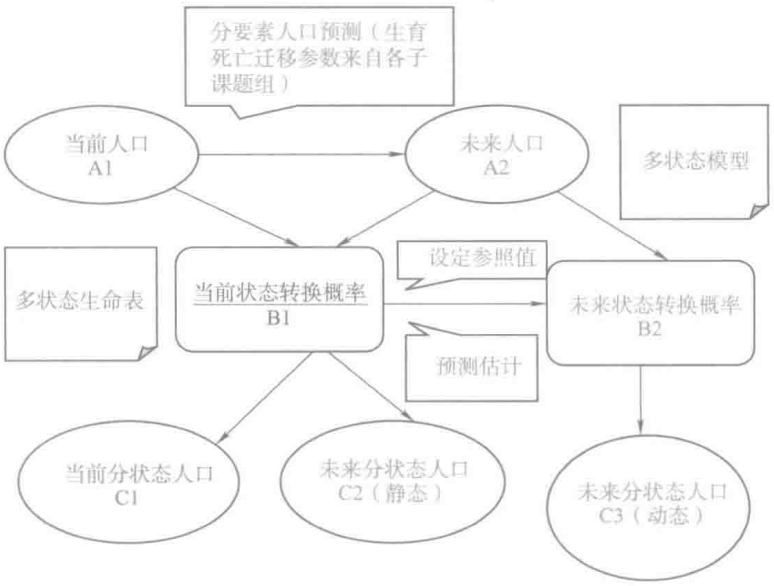


图 1　多状态人口分析预测关键技术与模型研究框架

说明:A1、B1、C1 可完成多状态生命表,与其他部分结合构成多状态模型,用于中长期预测。

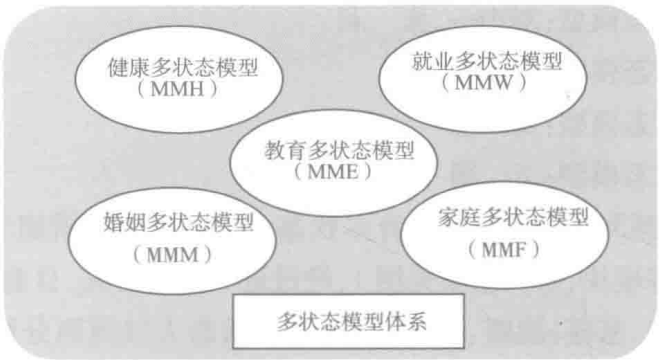


图 2　多状态人口分析预测关键技术与模型体系

目　录

第一章　多状态人口预测模型的数学基础

多状态人口学是处理同一人群在不同状态间转换的研究方法，使得人们能从生命历程视角更好地理解人口动态过程。多状态人口预测是对传统单递减生命表预测方法的拓展，其实质是多状态生命表，基础是马尔可夫过程，核心为转移概率矩阵，描述一个人口队列在婚姻变化、区域迁移、教育阶段、就业经历、健康变化等不同状态的进出和转移。本章将对多状态人口预测的数学基础进行介绍。

一、马尔可夫过程(Markov process)

马尔可夫过程是一类随机过程。在已知它目前的状态(现在)的条件下，它未来的演变(将来)不依赖于它以往的演变(过去)。这种已知“现在”的条件下，“将来”与“过去”独立的特性称为马尔可夫性，具有这种性质的随机过程叫作马尔可夫过程。液体中微粒所作的布朗运动、原子核中自由电子在电子层中的跳跃、传染病受感染的人数、森林中动物头数的变化构成、车站的候车人数、人口增长过程等都可视为马尔可夫过程。

二、马尔可夫链(Markov chain)

(一)离散时间马尔可夫链

1. 定义及性质

马尔可夫过程的原始模型为马尔可夫链，也称为离散时间马尔可夫链，即具有离散时间参数和离散状态空间的马尔可夫过程。对马尔可夫链，给定过去的状态 X_0, X_1, X_{n-1} 及现在的状态 X_n，将来的状态 X_{n+1} 的条件分布只依赖于现在的状态，而与过去的状态独立，称为马尔可夫性。该性质具有两个重要含义。第一，下一次转移的时刻不依赖于当前状态维持的时间长度，即在特定状态下的时间的分布具有“无记忆性”或“无后效性”。第二，下一个状态的概率分布仅仅依赖于当前的状态，这个过程所经历的状态序列构成一个马尔科夫链。

2. 转移概率

马尔可夫链的数学表达为如下随机过程：考虑只取有限个值的随机过程 $\{X_n, n =$

0,1,2…},可能取值的集合将以非负整数集{0,1,2…}来表示,即状态空间。若 $X_n = i$,就说过程在时刻 n 处于状态 i,假设每当过程处于状态 i,则在下一时刻将处于状态 j 的概率为一步转移概率 P_{ij}。即假设对一切状态 $i_0, i_1, \cdots, i_{n-1}, i, j$ 及一切 $n \geqslant 0$ 有

$$P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \cdots, X_1 = i_1, X_0 = i_0\} = P_{ij} \tag{1.1}$$

马尔可夫链的 n 步转移概率 P_{ij}^n 可定义为处于状态 i 的过程经 n 次转移后处于状态 j 的概率。即

$$P_{ij}^n = P\{X_{n+m} = j | X_m = i\}, n \geqslant 0, i, j \geqslant 0 \tag{1.2}$$

其算法由查普曼 - 柯尔莫哥洛夫(Chapman - Kolmogorov)方程,简称 C - K 方程提供。方程为:对于一切 $n, m \geqslant 0$,一切 i, j,有

$$P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^n P_{kj}^m \tag{1.3}$$

3. 马尔可夫链状态的分类

若对于某个 $n \geqslant 0$ 有 $P_{ij}^n > 0$,则称从状态 i 可到达状态 j。两个相互可到达的状态 i 与 j 称为相通的,记为 $i \leftrightarrow j$。相通的两个状态属于同一类,任意两个类或者不相交或者相同。

如果一个马尔可夫链只存在一个类,即一切状态彼此相通,则称该马尔可夫链是不可约的。

若从状态 i 出发经有限次转移后回到状态 i 的概率为 1,则称状态 i 是常返的,否则成为滑过的或非常返的。

假设状态有周期 d,满足 $d = G.C.D\{n: P_{ii}^n > 0\}$ (G. C. D:greatest common divisor,最大公约数),如果 $d > 1$,就称 i 为周期的,如果 $d = 1$,就称 i 为非周期的。

4. 转移概率矩阵(Transition Matrix)

马尔可夫在 20 世纪初提出,一个系统的某些因素在转移过程中,第 n 次结果只受第 $n-1$ 的结果影响,即只与当前所处状态有关,而与过去状态无关。他在分析中,引入状态转移这个概念。所谓状态是指客观事物可能出现或存在的状态;状态转移是指客观事物由一种状态转移到另一种状态。

如上,令 P_{ij} 代表处于状态 i 的过程下一步转移到状态 j 的概率,则所嵌入的马尔可夫链就具有转移矩阵 $\mathbf{P} = (P_{ij})$,即转移概率矩阵。

以 $\mathbf{P}$ 记一步转移概率 P_{ij} 的矩阵,从而

$$\mathbf{P} = \begin{bmatrix} P_{00} & P_{01} & P_{02} & \cdots \\ P_{10} & P_{11} & P_{12} & \cdots \\ \vdots & & & \\ P_{i0} & P_{i1} & P_{i2} & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}$$

由于概率是非负的，且过程必须转移到某个状态，所以有

$$P_{ij}\geqslant 0, i, j\geqslant 0; \sum_{j=0}^{\infty} P_{ij}=1, i=0,1,\cdots$$

举例来说，假设有如下马尔可夫链$\{X_n, n=0,1,2\}$，当$X_n=0$时，则$X_{n+1}=0,1$或2以相等的概率出现；当$X_n=1$时，则$X_{n+1}=0$以概率0.6出现，$X_{n+1}=1$以概率0.4出现；当$X_n=2$时，则$X_{n+1}=0$以概率1出现。则该马尔可夫链的一步转移概率矩阵$\mathbf{P}$有如下形式：

$$\mathbf{P}=\begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 0.6 & 0.4 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

以更为直观的状态转移图来表示，见图1－1。

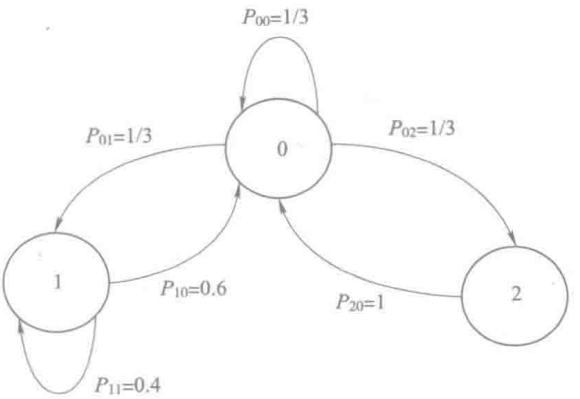


图1－1　状态转移图示例

以$\mathbf{P}^{(n)}$记n步转移概率P_{ij}^{n}的矩阵，则有$P^{(n+m)}=P^{(n)}\times P^{(m)}$，因此，

$$\mathbf{P}^{(n)}=P\times P^{(n-1)}=P\times P\times P^{(n-2)}=\cdots=\mathbf{P}^{n} \tag{1.4}$$

即n步转移矩阵$\mathbf{P}^{(n)}$是一步转移矩阵$\mathbf{P}$的n次方。

（二）连续时间马尔可夫链

1. 定义及基本概念

具有连续时间参数和离散状态空间的马尔可夫过程称为连续时间马尔可夫链。即设随机过程$\{X(t), t\geqslant 0\}$，状态空间$I=\{0,1,2\cdots\}$，若对任意$0\leqslant t_1<t_2<\cdots<t_{n+1}$及非负整数$i_1, i_2, \cdots, i_{n+1}$，有

$$\begin{aligned} &P\{X(t_{n+1})=i_{n+1}\mid X(t_1)=i_1, X(t_2)=i_2, X(t_n)=i_n\} \\ &\quad=P\{X(t_{n+1})=i_{n+1}\mid X(t_n)=i_n\} \end{aligned} \tag{1.5}$$

则称$\{X(t), t\geqslant 0\}$为连续时间马尔可夫链。

2. 转移概率

在已知现在s时刻的状态及一切过去状态的条件下，其在将来时刻$t+s$的状

态的条件分布只依赖于现在的状态而与过去独立。其 s 时刻处于状态 i,经过时间 t 后转移到状态 j 的转移概率为

$$P_{ij}(s,t)=P\{X(s+t)=j|X(s)=i\} \tag{1.6}$$

其转移概率矩阵简记为

$$\mathbf{P}(\mathbf{s},\mathbf{t})=[P_{ij}(s,t)] \tag{1.7}$$

若 $\{X(t),t\geqslant 0\}$ 的转移概率与 s 无关,即 $P_{ij}(s,t)=P\{X(s+t)=j|X(s)=i\}=P_{ij}(t)$,则称连续时间马尔可夫链具有平稳的或齐次的转移概率,为齐次连续时间马尔可夫链。其转移概率矩阵简记为

$$\mathbf{P}(\mathbf{t})=[P_{ij}(t)] \tag{1.8}$$

马尔科夫链初始时刻各状态的概率 $P_i=P_i(0)=P\{X(0)=i\}$,$i\in I$ 为初始概率,其分布 $\{P_i,i\in I\}$ 为初始分布;在时刻 t 的 $P_i(t)=P\{X(t)=i\}$,$i\in I$ 为绝对概率,其分布 $\{P_i(t),i\in I\}$ 为绝对分布。

对状态有限的马尔科夫链,如果存在 $k>0$,使 $P_{ij}(k)>0,i,j=1,2,\cdots,N$,则称此马尔科夫链具有遍历性,且具有极限分布 $\pi=(\pi_1,\pi_2,\cdots,\pi_N)$,它是方程组 $\pi=\pi\mathbf{P}$,或 $\pi_i=\sum_{i=1}^{N}\pi_iP_{ij},j=1,2,\cdots,N$ 的满足条件 $\pi_i>0$,$\sum_{i=1}^{N}\pi_i=1$ 的唯一解,称 $\pi=(\pi_1,\pi_2,\cdots,\pi_N)$ 为平稳分布。

三、马尔可夫链在人口学中的应用

连续时间马尔可夫链的一类重要的特殊情形是生灭过程,其特征是在很短的时间内,系统的状态只能从状态 i 转移到状态 $i-1$ 或 $i+1$ 或保持不变。其确切定义是:设齐次连续时间马尔可夫链 $\{X(t),t\geqslant 0\}$ 的状态空间是 $I=\{0,1,2,\cdots\}$,转移概率为 $P_{ij}(t)$,如果满足如下条件:

$$\begin{cases}P_{i,i+1}(h)=\lambda_i h+o(h),\lambda_i>0;\\ P_{i,i-1}(h)=\mu_i h+o(h),\mu_i>0,\mu_0=0;\\ P_{ii}(h)=1-(\lambda_i+\mu_i)h+o(h);\\ P_{ij}(h)=o(h),|i-j|\geqslant 2\end{cases}$$

则称 $\{X(t),t\geqslant 0\}$ 为生灭过程。其中,λ_i 为出生率,若 $\lambda_i=0$ 则为纯灭过程;μ_i 为死亡率,若 $\mu_i\equiv 0$ 则为纯生过程。若 $\lambda_i=i\lambda,\mu_i=i\mu$($\lambda,\mu$ 正常数),则称 $\{X(t),t\geqslant 0\}$ 为线性生灭过程。

在人口过程中的生育和死亡即符合生灭过程。如果以 $X(t)$ 表示一个人群在 t 时刻的大小,则在很短的时间 h 内(忽略其高阶无穷小量 $o(h)$),群体变化有三种可能:

状态由 i 变到 $i+1$,即增加一个个体,其概率为 $\lambda_i h$;

状态由 i 变到 $i-1$,即减少一个个体,其概率为 $\mu_i h$;

群体大小不增不减,其概率为 $1-(\lambda_i+\mu_i)h$。

此外，人口过程还包括婚姻、家庭、教育、就业、健康等多种不同的状态过程。以婚姻为例，从未婚的初始状态，可能转移到已婚，从已婚又可转移为离婚、丧偶，离婚又可转移为再婚，以此类推，未来的状态只依赖于现在的状态而与过去无关。类似的，就业状态、家庭类型状态、教育状态、健康状态等的转换过程都是符合马尔可夫性的马尔可夫过程。其中婚姻、就业、家庭类型和健康状态往往具有连续的时间参数和离散的状态空间，因此可归为连续时间马尔可夫链，而教育状态则具有较为明显的离散时间参数和空间状态，因此可归为离散时间马尔可夫链。更为详细的各状态模型将在后续章节分别进行介绍。

第二章　教育多状态模型的理论、模型与实践

第一节　理论综述

教育多状态人口预测是对传统人口预测方法的扩展，根据年龄、性别和教育程度来分析某一区域的人口在预测期内受死亡、人口净迁移影响而变化。队列内不同教育程度人数则在小学、中学和大学等适龄阶段随着教育转换发生变化。

一、技术路线

（一）模型方法

在过去十年里，国际应用系统分析研究所（IIASA）在“按教育程度分的多状态人口预测”中深入研究不同教育水平的人口会有不同的生育和死亡水平的重要影响，将其作为教育多状态参考模型。

国际应用系统分析研究所（IIASA）在 1970～1980 年的人口项目，参加者有：Andrei Rogers，Frans Willekens 和 Jacques Ledent。在过去十年里，IIASA 多状态预测方法明显拓展和应用是“按受教育程度分的多状态人口预测（MPPEA）”。从 Luts 和 Goujon（2001：325）论文摘出的图 2－1 显示，MPPEA 明确地指出了人口在不同教育水平上会有不同的生育率和死亡率水平这一重要事实。受教育程度预测基于两个不同的人群：对那些仍然停留在受教育过程中的人，该模型预测了按照年龄和性别分的从一个教育水平到另一个较高的教育水平的转换。对那些已经完成他们最高教育水平的人，该预测简单地随着时间增加他们的年龄，并按照教育别死亡率、生育率和迁移率将他们计算出来。MPPEA 模型直接地和清晰地演示了从学校到人力资本的转换过程，采用整个劳动力的平均教育水平进行度量。

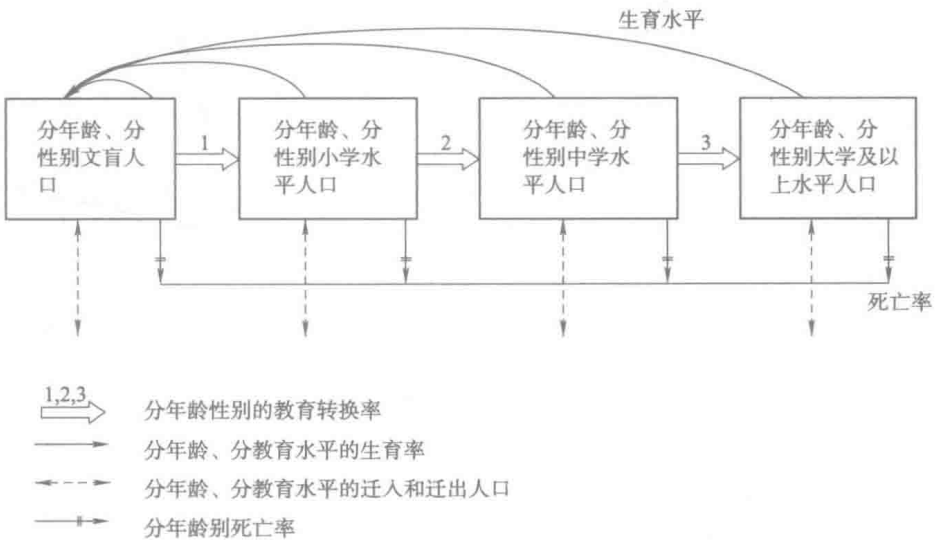


图 2-1　按教育水平分的多状态人口预测模型结构

资料来源：Wolfgang Lutz and Anne Goujon (2001). The world's changing human capital stock: multi-state population projection by educational attainment. Population and Development Review, 27(2): 325.

(二)模型关键参数

1. 基年人口数据(分年龄、性别、城乡、教育程度)。

2. 生育率。

3. 死亡率。

4. 教育转换率。

5. 迁移率。

其中,教育转换率是关键参数,指某一教育程度人口中在预测期内获得更高一级教育程度的人口的比例(6~24 岁,5 岁一组)。

第一个年龄组的教育转换率：$T_{ij}(AgeG_n) = \dfrac{popr_i(AgeG_n)}{popr_i(AgeG_{n-1})}$

第二个年龄组的教育转换率：$T_{ij}(AgeG_{n+1}) = 1 - \dfrac{popr_i(AgeG_{n+1})}{popr_i(AgeG_n)}$

$T_{i,j}$是从状态 i 到状态 j 的转换率,$popr_i$ 是状态 i 的人口占该年龄组总人口的比率。

文盲到小学的转换率如下：

5~9 岁组文盲到小学的转换率：$T_{1,2}(5\sim9) = \dfrac{popr_2(5\sim9)}{popr_1(0\sim4)}$

10~14 岁组文盲到小学的转换率：$T_{1,2}(10\sim14) = \dfrac{popr_2(10\sim14)}{popr_1(5\sim9)}$

其他各个相应的转换率如下：