State Prediction of Policy-Holder of Basic Endowment Insurance for Urban Employees Based on Markov Chain

Tianyang Lv [1], Yuhui Qiu [2], Shaobin Huang [1], Qi Pang [1]

[1] *College of Computer Science and Technology, Harbin Engineering University, Harbin, 150001*

[2] *College of Philosophy and Public Administration, Heilongjiang University, Harbin, 150080*

Abstract: Basic endowment insurance plays an important role in the social insurance system. At present, the researches on basic endowment insurance are mainly qualitative analysis which comes from the commercial actuarial methods dealing with a few affirmatory state changes of policy-holders. And the quantitative researches are relatively few. However, the state changes in social insurance are uncertain. After analysing the policy and the characteristics of basic endowment insurance of China, the paper proposes a novel method to describe the policy-holder's state of basic endowment insurance for urban employees. The method has the no-after-effect property. Therefore, the Markov chain is adopted to forecast the state change of policy-holders based on the state transition matrix. Finally, experiments are performed based on the real data of one million policy-holders of two different areas. And the experimental result shows the good performance of our method.

Keywords: endowment insurance; social insurance actuarial; Markov chain; forecast

基于马尔可夫链的基本养老保险参保人员状态预测

吕天阳 [1]，邱玉慧 [2]，黄少滨 [1]，庞琦 [1]

[1] 哈尔滨工程大学 计算机科学与技术学院 哈尔滨 150001

[2] 黑龙江大学 哲学与公共管理学院 哈尔滨 150080

摘　要：基本养老保险是我国社会保险体系的重要组成部分。目前对基本养老保险的定量研究方法主要基于统计数据，采用由商业养老保险发展而来的精算方法。但是与商业养老保险相比，城镇企业职工基本养老保险参保人员的参保状态变化更为复杂且具有明显的不确定性。本文通过分析基本养老保险相关政策及业务特点，采用一种新型的具有无后效性的参保人员状态描述方法，建立参保人员状态迁移矩阵，从而应用马尔可夫链预测参保状态的变化和供养比的发展趋势。最后，使用两个统筹区域的海量真实的基本养老保险数据开展实验，验证方法的有效性。

关键字：基本养老保险，社会保险精算，马尔可夫链，预测

1 背景

社会保险是由国家立法规范的，面向劳动者建立的一种强制性社会保障制度，包括养老保险、医疗保险、失业保险、工伤保险、生育保险等项目[1]。其目的在于，通过社会集资，保证在因不可避免风险而暂时或永久失去劳动收入后，劳动者能够获得一定程度的收入补偿[2]。

我国养老保险制度包括城镇企业职工基本养老保险、机关事业养老保险和农村养老保险。本文的研究重点为城镇企业职工基本养老保险（以下简称为基本养老保险）。90 年代以来，国家出台了一系列有关基本养老保险的规定，界定了个人、企业和政府的三方权责，采取部分积累制下的分账户法，奠定了制度的框架。截至 2009 年底，基本养老保险参保人数达到 2.3 亿人，基金积累已达万亿。

基本养老保险基金的可持续性是我国养老保险的重要研究课题。目前，国内研究者主要采用定性方法研究基本养老保险的可持续性，且主要借用商业保险中的精算方法[3-4]；定量研究相对较少，而且主要以统计性数据作为数据基础，其准确性有待商榷。

本文受国家科技支撑计划 2009BAH42B02，国家自然科学基金 60873038，黑龙江省哲学社会科学研究项目 08E061，中央高校基本科研业务专项资金项目 HEUCF100603 资助。本文部分研究成果为第一作者在审计署国家审计数据中心研究期间取得。

与商业保险相比，基本养老保险参保人员状态变化更为复杂且具有明显的不确定性。例如，由于特殊工种年限、国有企业破产、地方政策、违规操作等等因素，参保人员究竟在何年龄退休，是一个不确定的问题。一些研究者假设参保人员的退休年龄为一个常量，例如 58 岁[5]，并不符合实际情况，必然影响预测结果的有效性。

为此，本文采用海量真实的基本养老保险数据，基于马尔可夫链预测参保人员的状态变化。通过分析基本养老保险相关政策及业务特点，本文提出了能够描述基本养老保险体制下参保人员状态迁移的模型，使得参保人员的状态变化具有无后效性，并解决了状态空间爆炸的问题。基于该模型，建立参保人员状态迁移矩阵，从而应用马尔可夫链预测参保状态变化的趋势，进而预测基本养老保险基金供养比的发展趋势。最后，使用两个区域的海量真实的基本养老数据开展实验，验证了方法的有效性。

与以往研究相比，本文更精确的描述了参保人员状态变化的不确定性，实验的数据基础不是抽样或统计数据，而是利用审计署审计数据中心的接近 1 百万参保人员的真实数据。因此，具有更强的真实性和可操作性。此外，据本文作者掌握，尚无研究应用马尔可夫链的方法解决社会保险研究中的问题，而且本文所采用的状态数（上百个）明显多于其他领域的状态数（一般不超过 10 个）[7, 8, 9, 10]。

本文其他部分的内容安排如下：第 2 节介绍参保人员的状态迁移模型，第 3 节给出无后效性的参保人员状态描述，第 4 节给出基于马尔可夫链的参保人员状态描述，第 5 节为实验与分析，第 6 节总结全文。

2 参保人员基本状态变化模型

由于在基本养老保险制度中，参保人员的状态变化具有不确定性。因此，需要首先分析参保人员的状态，才能给出基本养老保险所有可能的参保状态变化情况。

以是否参加基本养老保险制度为分界，城镇企业职工可分为三种基本状态：参保、未参保与死亡。参保人员可以维持参保状态，也可以通过退保转为未参保状态或因为死亡变为死亡状态；未参保人员可以维持未参保状态，也可以通过参加基本养老保险转为参保状态或因为死亡变为死亡状态；处于死亡状态的人员其状态将不再发生变化。

同商业养老保险相比，基本养老保险参保人员除有参保缴费与退休享受待遇两种状态之外，还具有两种状态：参保未缴费和转移（包括统筹区域间转移和区域内转移）。参保缴费人员可以维持缴费状态，也可能由于暂停缴费转为参保未缴费状态或变为退休享受待遇状态；参保未缴费人员可以维持参保未缴费状态，也可以恢复缴费或变为退休享受待遇状态。

至此，在考虑参保人员关系转移的情况下，基本养老保险制度参保人员共计 6 个状态：未参保、参保未缴费、参保缴费、转移、退休领取待遇和死亡。其中，参保人员重要的状态变化，共计 21 种，参见图 1。其中，死亡状态为吸收态。

进一步分析基本养老保险业务，当参保人员处于参保缴费或参保未缴费状态时，参保人员退保或死亡时，基本养老保险的处理方式基本相同，且未参保人员死亡与基本养老保险无关，因此人员状态变化关系可以进行化简。同时由于很难掌握参保人员转移的具体方向，因此，也可以暂不考虑转移状态的影响。化简后的状态变化关系见图 2。

3 无后效性的参保人员状态描述

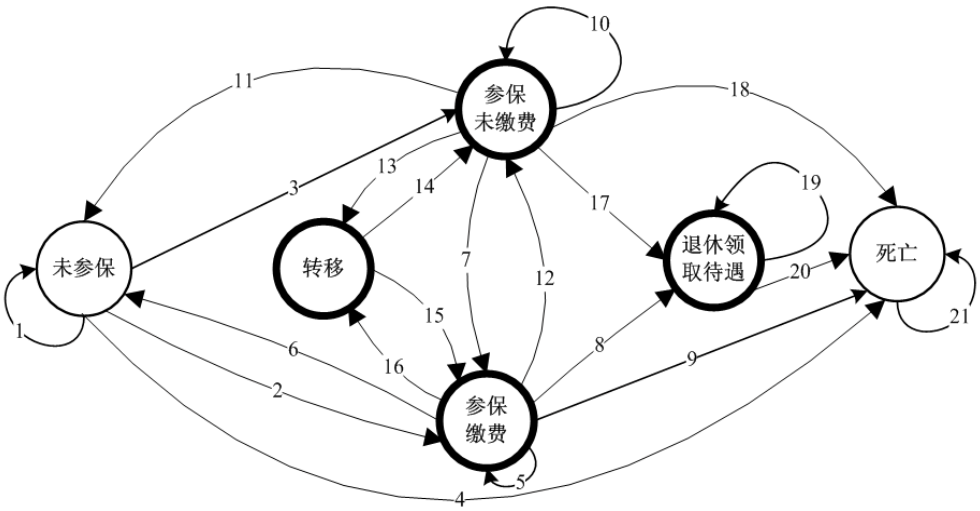
[1] 第 2 节给出了参保人员的状态迁移图，其中，由参保缴费状态向享受待遇状态的变迁，对基本养老保险基金的平衡性冲击最大，因此是研究的重点。一种解决方式是：寻找具有无后效性的参保人员状态描述，使得参保人员的历史因素体现在当前的状态中，从而在预测中不必考虑此前的状态。无后效性是马尔可夫链的重要属性，是指“过程或（系统）在时刻 t_0 所处的状态为已知的条件下，过程在时刻 $t > t_0$ 所处状态的条件分布，与过程在时刻 t_0 之前所处的状态无关”[6]。

参保状态外还需要考虑性别、年龄与缴费年限三种因素。因此，通过将这四种属性组合的方式描述参保人员状态。

令 $X^t_{C,\ S,\ A,\ Y}$ 表示参保人员在 t 时刻的状态，其中 C 表示参保人员的参保状态集合{j=参保缴费状态，u=参保未缴费状态，r=退休状态}；S 表示参保人员的

978-1-935068-18-1 © 2010 SciRes.

性别集合{*m*=男性, *f*=女性}；*A* 表示参保人员的年龄集合{18, 19.....100}，即 18 岁-100 岁；*Y* 表示参保人员的缴费年限集合{1, 2.....42}，即最多缴费 42 年。



1 人员仍处于未参保状态；2 未参保人员参加基本养老保险并缴费；3 未参保人员参加基本养老保险但未缴费；4 未参保人员死亡；5 参保人员继续参保缴费；6 参保人员停止缴费并解除基本养老保险关系；7 参保人员停止缴费但未解除基本养老保险关系；8 参保人员退休领取待遇；9 在职缴费人员死亡；10 参保未缴费状态未发生变化；11 参保未缴费人员解除基本养老保险关系；12 参保未缴费人员恢复缴费；13 参保未缴费人员转出；14 参保未缴费人员转入；15 参保缴费人员转入；16 参保缴费人员转出；17 参保未缴费人员退休领取待遇；18 参保未缴费人员死亡；19 退休领取待遇人员状态未发生变化；20 退休领取待遇人员死亡；21 死亡状态吸收态。

Figure 1. State transition graph of policy-holder of basic endowment Insurance for urban employees
图 1 基本养老保险参保人员状态迁移图

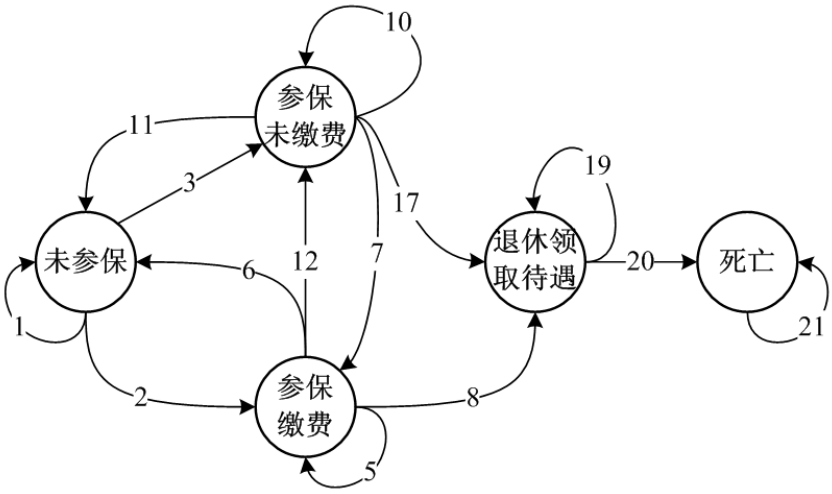


Figure 2. Simplified state transition graph of policy-holder of basic endowment Insurance
图 2 简化的参保人员状态迁移图

在该方法的描述下，参保人员当前状态的信息中已经记录了影响参保人员的历史信息，所以各类参保人员的状态变化具备无后效性，因此每类参保人员的状态变化过程拥有马尔可夫链的相关性质。例如，某参保人员处于 $X^{t}_{j,m,a,y}$ 状态，当 a 与 y 的值满足退休条件时，该人员在 t+1 时刻参保状态可能的变化情况如下：（1）继续参保缴费，状态变为 $X^{t+1}{}_{j,m,a,y}$；（2）停止缴费，状态变为 $X^{t+1}{}_{u,m,a,y}$；（3）退休享受待遇，状

态变为 $X^{t+1}{}_{r,m,a,y}$；（4）死亡，吸收态记为 X_D，参见图 3。

由于采用了性别、年龄与缴费年限补充参保人员的状态描述，使得状态空间爆炸。分析如下：令 n_C 表示参保状态属性的可取值数；n_S 表示性别属性的可取值数；n_A 表示年龄属性的可取值数；n_Y 表示缴费年限属性的可取值数。则状态总数为：

$$N = N_C \times N_S \times N_A \times N_Y = 3\times 2\times 83\times 42 = 20916 \quad (1)$$

这一状态数已经超过了实际数据的支撑能力。由于基本养老保险基本处于市级统筹，因此统筹地区的参保人数达到 100 万已是人数较多。此时，平均约有 50 个参保人员处于一个参保状态，而且存在部分状态中只有几个参保人员甚至没有参保人员的情况，从而影响到参保人员状态变化预测结果的可信度。

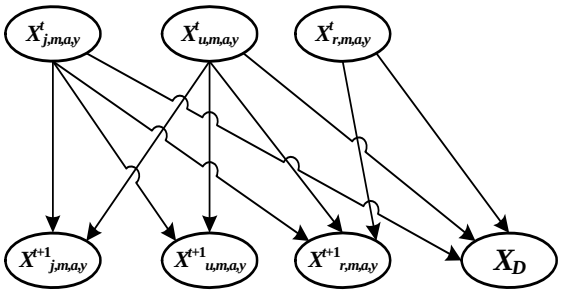


Figure 3. Part of state changes of policy-holder

图 3 参保人员状态部分变化情况

为此，需要简约状态。一方面，在牺牲一定精度的情况下，合并状态，例如将缴费年限和年龄视为一种因素，共同影响退休条件。另一方面，剔除无法满足的状态，例如，男性参保人员正常缴费的年龄应该在 18 岁至 60 岁之间，离退休享受待遇年龄在 50 岁至 100 岁。男性参保未缴费人员的情况类似。则男性参保人员的状态总数 Nm 通过公式（2）计算：

$$N_m = 43 + 43 + 51 = 137 \quad (2)$$

同理可得，女性参保人员的状态总数 N_f 如式（3）所示：

$$N_f = 38 + 38 + 56 = 132 \quad (3)$$

可见，约简后参保人员的状态总数减少为 N = 269。

4 基于马尔可夫链的参保人员状态预测模型

定义 1：称状态空间 E = {X_D, X_W, $X_{C,S,A}$}为基本养老保险人员状态空间。

其中 X_W 表示未参加基本养老保险的状态。

定义 2：迁移概率 $P\{X_{C',S',A'} \mid X_{C,S,A}\}$ 表示状态为 $X_{C,S,A}$ 的参保人员其参保状态变化至 $X_{C',S',A'}$ 的概率。迁移概率也记为 $P_{(C,S,A),(C',S',A')}$。

当样本数据充足时，则可通过公式（4）计算状态变化概率

$$p_{(C,S,A),(C',S',A')} = \frac{f(X_{(C,S,A),(C',S',A')})}{f(X_{C,S,A})} \quad (4)$$

其中，$f(X_{C,S,A})$ 表示样本数据中状态 $X_{C,S,A}$ 出现的次数；则 $f(X_{(C',S',A'),(C,S,A)})$ 表示样本数据中参保人状态由 $X_{C,S,A}$ 变化至 $X_{C',S',A'}$ 的总次数。

在状态空间 ***E*** 中，死亡状态 X_D 为吸收态，由于处于 X_W 状态的人数无法通过社会保险数据统计，所以无法直接计算出与之相关的状态变化概率，因此假定新参保人员参保后平均分布在与缴费相关的各个状态，即：

$$P\{X_{j,S,A} \mid X_W\} = \frac{1}{81} \quad (5)$$

由于未参保人员死亡或维持状态不变对基金没有任何影响，所以不妨设 $P\{X_D \mid X_W\} = P\{X_W \mid X_W\} = 0$，则有：

定义 3：状态迁移矩阵 M

$$P = \begin{array}{c} \\ X_D \\ X_W \\ X_{j,m,18} \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} X_D & X_W & X_{j,m,18} & \cdots \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & \cdots \\ 0 & 0 & 1/81 & \cdots \\ p_{(j,m,18),D} & p_{(j,m,18),W} & p_{(j,m,18),(j,m,18)} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{array} \quad (6)$$

矩阵中任意元素 P_{ij} 的取值为由状态 i 迁移到状态 j 的概率 $P\{j|i\}$，同时满足：$\sum_j P_{ij} = 1$ 且 $P_{ij} \le 1$。

基于 P 可以预测参保人员状态迁移情况，并给出各时间点处于相应状态的参保人员的人数。不妨设，t 时刻处于各参保状态的人数为行向量 $N_t = \{n_D^t, n_W^t, n_{(j,m,18)}^t \ldots\ldots\}$。则（$t$+1）时刻处于各参保状态的人数为：

$$N_{t+1} = N_t M + N_W^{t+1} \quad (7)$$

其中 n_W^t 为 t 时刻新增参保人数，$N_W^{t+1} = \{0, n_W^{t+1}, 0 \ldots\ldots\}^T$。

978-1-935068-18-1 © 2010 SciRes.

至此，可以利用大量基本养老保险业务数据预测参保人员状态变化。

5 实验与分析

实验采用审计署数据中心两个地区真实的基本养老保险数据。其中，B 地区 2005 年至 2008 年的参保人数分别为 74.6 万、86.6 万、64.6 万和 55.5 万。C 地区参保人数在 10 万左右。

5.1 B 地区参保人员状态变化预测

根据计算出的各状态之间变化的概率结果，可以得到 B 地区人员状态变化矩阵。

$$P=\begin{array}{c}\\ X_D\\ X_W\\ X_{j,m,18}\\ X_{j,m,19}\\ X_{j,m,20}\\ X_{j,m,21}\\ X_{j,m,22}\\ X_{j,m,23}\\ \vdots\end{array}\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0.012 & 0.012 & 0.012 & 0.012 & 0.012 & 0.012 & \cdots \\ 0 & 0.114 & 0 & 0.886 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0.234 & 0 & 0 & 0.766 & 0 & 0 & 0 & \cdots \\ 0 & 0.262 & 0 & 0 & 0 & 0.738 & 0 & 0 & \cdots \\ 0 & 0.285 & 0 & 0 & 0 & 0 & 0.715 & 0 & \cdots \\ 0 & 0.298 & 0 & 0 & 0 & 0 & 0 & 0.702 & \cdots \\ 0 & 0.320 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (8)$$

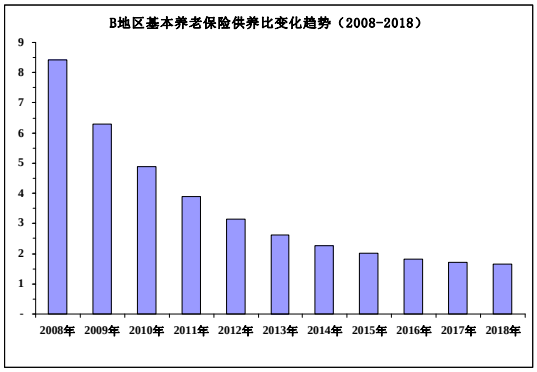


Figure 4.the trend of support rate of basic endowment insurance in area B (2008-2018)

图 4 B 地区基本养老保险供养比变化趋势（2008-2018）

以预测出该地区未来十年的基本养老保险参保人员状态变化情况。预测结果表明，该地区未来十年内缴费人数大大减少，而享受待遇人数基本保持稳定。根据 B 地区的基本养老保险参保情况变化可以计算出基本养老保险供养比（即缴费人数与享受待遇人数之比）在未来十年内逐年下降，具体变化情况见图 4。根据 B 地区 2008 年的参保情况及收支情况，可计算出 2008 年该地区年平均缴费金额为 5942.53 元，年平均待遇金额为 12470.74 元。进而计算出在当前缴费率与替代率情况下，供养比应维持在 2.10 以上才能保证该地区基本养老保险基金不出现收支赤字。因此未来几年内该地区在基本养老保险缴费率与替代率不变的情况下仍能保持收支结余；但从 2015 年起，该地区基本养老保险基金将开始现收不抵支的情况。

5.2 C 地区基本养老保险基金趋势预测

本文再以 C 地区为例，对其基本养老保险基金的发展趋势进行预测。对其基本养老保险数据进行与 B 地区类似的处理方式。

可以计算出 2008 年供养比的变化趋势，具体情况见图 5。而 C 地区 2008 年的年平均缴费金额为 2770.80 元，年平均待遇金额为 9216.66 元，由此可计算出在该地区的缴费率与替代率不发生变化的情况下，其供养比应维持在 3.33 以上才能保证该地区的基本养老保险基金不出现收支赤字。可以看出 C 地区的基本养老保险基金已经处于收不抵支的状态中，并且若不加紧调整相关政策，未来十年间该地区的基本养老保险基金的收支缺口将逐年扩大。同时可以看出虽然 B 地区与 C 地区的基本养老保险基金都因供养比的逐年降低而收支压力加大，但其根本原因缺有区别，B 地区主要是因为缴费人数的减少，而 C 地区主要是因为享受待遇人数的增加。

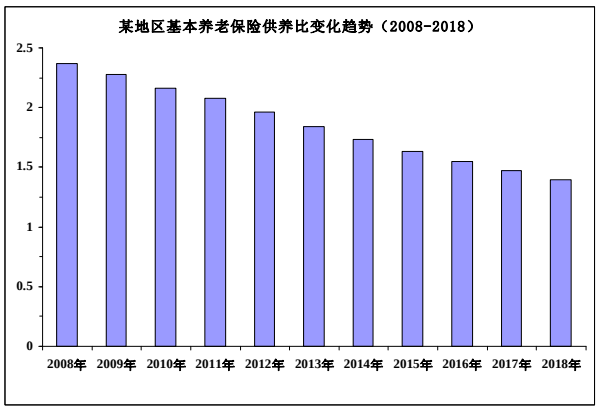


Figure 5. the trend of support rate of basic endowment insurance in area C (2008-2018)

图 5 C 地区基本养老保险供养比变化趋势（2008-2018）

通过对 B 地区与 C 地区基本养老保险基金发展趋势的预测，表明本文所使用的预测方法可以预测出基本养老保险基金的发展趋势，并且对业务数据的加工

处理既能够体现不同地区基本养老保险的不同特点，又避免了大量业务数据不便存储与迁移的问题。

5.3 预测结果比较

分别使用基本养老保险精算方法与本文所采用的方法对 2008 年的参保人员状态分布情况，而后与实际情况进行比较分析。比较结果见表 1。

根据预测结果与实际情况的比较，可以发现基本养老保险精算中，对参保人员退休年龄的估计过于乐观，从而导致在揭示基本养老保险基金将要面对的问题是，一定程度的掩盖了问题的规模。

Table 1. Result comparison of the proposed method and actuarial method
表 1 本文方法与精算方法预测结果

	本文方法		精算方法		实际值
	预测值	相对误差	预测值	相对误差	
参保总人数（人）	523877	-5.56%	523877	-5.56%	554725
缴费人数（人）	487721	-1.62%	492483	-0.66%	495757
缴费人员所占比例	93.10%	4.17%	94.01%	5.29%	89.37%

6 总结

本文针对基本养老保险参保人员参保状态变化复杂的情况，进行了以下主要工作：分析了基本养老保险相关政策及业务特点，采用了一种基本养老保险特点的参保人员状态描述方法，并以此为基础，通过预测参保人员状态变化情况，对基本养老保险基金发展趋势进行预测。使用实际数据检验预测方法，并将本文所使用方法与精算方法的预测结果进行比较。实验表明，本文所使用方法的预测结果更为准确。

下一步的工作是：将该方法推广到其他社会保险；分析新的社会保险政策的影响，尤其是 2009 年参保人员转移接续办法的影响。

References (参考文献)

[1] Zheng Gongcheng. Discipline of Social Security. China Labour and Social Security Press, 2005.7.
郑功成. 社会保障学. 中国劳动社会保障出版社, 2005 年 7 月.

[2] Hou Wenruo. Social Insurance. China Labour and Social Security Press, 2005.7.
侯文若. 社会保险. 中国劳动社会保障出版社, 2005 年 7 月.

[3] Yu Hong, Zhong Heqing. On the Sustainable Operation of China's Basic Endowment Insurance System. Journal of Finance and Economics, 2009(9), 26-35P.
于洪，钟和卿. 中国基本养老保险制度可持续运行能力分析. 《财经研究》(沪). 2009 年 9 月. 26~35.

[4] Xu Ying, Wang Jianwei. Evaluation and Analysis of Designed Substitution Rate of Basic Endowment Insurance System in Urban China. Population & Economics, 2009(4), 78-84P.
徐颖，王建梅. 对城镇基本养老保险制度设计替代率的评估分析. 《人口与经济》(京). 2009 年 4 月. 78~84.

[5] Zhang Sifeng. Principles and Application of Social Security Actuarial. People's Publishing House. 2006.
张思锋. 社会保障精算理论与应用. 北京：人民出版社, 2006.

[6] Encyclopaedia of mathematics. Vol. 3. Science Press. 1997
数学百科全书.(第三卷). 科学出版社, 1997 年.

[7] Qian jiazhong, Zhu Xuesi, Wufengsi. Time Series-Markov Prediction Model for Precipitation in the Course of Evaluation of Groundwater Resources. Scientia Geographica Sinica, 2001,12(4).
钱家忠，朱学愚，吴剑锋. 地下水资源评价中降水量的时间序列马尔可夫模型. 地理科学, 2001,12(4).

[8] Sun Caizhi etc.. Research on fuzzy Markov chain model with weights and its application in predicting the precipitation state. Journal of Systems Engineering, 2003,18(4).
孙才志，林雪钰. 降水预测的模糊权马尔可夫模型及应用. 系统工程学报, 2003,18(4).

[9] Butt A A, Shahin W Y, Feighan K J et al.Pavement performance prediction model using the Markov process.Transportation Research Record, 1987: 12-19P

[10] Yu Xunqi, James W Modestino, Tian Xusheng. The accuracy of Markov chain models in predicting packet-loss statistics for a single multiplexer. IEEE Transactions on Information Theory, 2008(1): 489-501P.

978-1-935068-18-1 © 2010 SciRes.