中国人口增长预测

摘　　要

本文针对中国人口增长预测问题，建立了基于 Gompertz 模型的人口增长模型、基于灰色预测模型的出生人口性别比预测模型、基于 Leslie 修正模型的城镇化预测模型、基于神经网络的老龄化趋势预测模型和基于模糊线性回归分析的人口预测模型，分别解决了中国人口数量的大致趋势预测、出生人口性别比预测、中国城镇化预测、老龄化趋势预测以及中国人口数量更精准的预测问题。

针对中国人口数量的大致趋势预测问题，建立了于 Gompertz 模型的人口增长模型。首先推导出 Gompertz 增长基本规律的数学公式并根据我国的实际情况对公式变形，得到人口增长模型；其次查找相关数据并通过最小二乘法拟合出公式的系数；最后预测出中国人口数量的增长趋势。最终得到 2006-2050 年的中国人口总数呈上升趋势，每年的增长人数在 0.05-0.2 亿人之间；至 2050 年，我国的人口数量约为 15.4154 亿。

针对出生人口性别比预测问题，建立了基于灰色预测模型的出生人口性别比预测模型。首先整理得到我国 1995-2005 年市、镇、乡的出生人口性别比的数据。其次建立灰色预测模型预测我国 2006-2050 年我国市、镇、乡的出生人口性别比。最后通过残差和级比偏差检验所建模型的准确性。最终得到 2006-2050 年我国未来城、镇、乡的出生人口性别比呈缓慢增长的趋势。

针对中国城镇化预测问题，建立了基于 Leslie 修正模型的城镇化预测模型。首先对我国人口数据进行初步处理分析;其次建立 Leslie 修正模型预测出中国 2006-2050 年的城镇人口数。其次通过城镇人口数和总人口数计算城镇化率；最后通过城镇化率分析中国未来城镇化问题的走向。最终得到 2035 年以后中国城镇化率将达到 70%以上,到 2050 中国城镇化率将达到 77.0765%。

针对老龄化趋势预测问题，建立了基于神经网络的老龄化趋势预测模型。首先根据我国人口统计数据的特性将人口老龄化数据模糊化，建立老龄化模糊集；其次根据我国现状设计老龄化指标；最后建立基于神经网络的老龄化趋势预测模型对我国未来老龄化趋势进行预测。最终得到未来我国人口老龄化指标的变化趋势在 2010 前之前有较大幅度的增长，2010 年之后发展较平稳，逐步稳定在 0.4730。

针对中国人口数量更精准的预测问题，建立了基于模糊线性回归分析的人口预测模型。首先将已知数据模糊化；其次根据以上三个痛点问题对我国人口增长模型的影响建立基于模糊线性回归分析的人口预测模型；最后利用模糊最小二乘法确定模型的系数，预测中国未来人口数量。最终预测出我国中短期、长期的人口数量，相比于基于 Gompertz 模型的人口预测模型，精确度有很大提高。

关键词：Gompertz 模型　灰色预测模型　Leslie 修正模型
神经网络　模糊线性回归分析模型

一、问题重述

中国是一个人口大国，人口问题始终是制约我国发展的关键因素之一。根据已有数据，运用数学建模的方法，对中国人口做出分析和预测是一个重要问题。

近年来中国的人口发展出现了一些新的特点，例如，老龄化进程加速、出生人口性别比持续升高，以及乡村人口城镇化等因素，这些都影响着中国人口的增长。2007 年初发布的《国家人口发展战略研究报告》(附录 1) 还做出了进一步的分析。

关于中国人口问题已有多方面的研究，并积累了大量数据资料。附录 2 就是从《中国人口统计年鉴》上收集到的部分数据。

试从中国的实际情况和人口增长的上述特点出发，参考附录 2 中的相关数据（也可以搜索相关文献和补充新的数据），建立中国人口增长的数学模型，并由此对中国人口增长的中短期和长期趋势做出预测；特别要指出你们模型中的优点与不足之处。

附录 1 《国家人口发展战略研究报告》

附录 2　人口数据（《中国人口统计年鉴》中的部分数据）及其说明

二、问题分析

首先，为了准确预测我国未来人口的增长数量，我们首先想到符合 S 型增长曲线的 Logisitic 模型。但是由于客观因素等不可抗力的存在，Gompertz 模型更符合人口增长的基本规律。因此，我们选择 Gompertz 模型建立我国未来人口预测的模型，计算出 2006-2050 年我国的人口总数。

其次，我们对我国痛点问题进行分析，即出生人口性别比持续升高、老龄化进程加速以及乡村人口城镇化问题。第一，对于预测未来的城、镇、乡的出生人口性别比，我们选择利用灰色预测，通过残差和级比偏差检验所建模型的准确性。第二，我们选用 Leslie 模型对我国未来的城镇化水平进行预测，通过分析近年来的总和生育率、死亡率，预测中国未来城镇化问题的走向。最后，针对我国的老龄化问题，我们选用神经网络模型进行预测。由于老龄化是一个不确切的指标，我们要先将人口老龄化问题魔模糊化，再计算老龄化指标，最后利用神经网络对我国未来老龄化趋势进行预测。

最后，对我国未来的总人口数进行预测。考虑到人口普查中的错报、漏报情况，我们选择利用模糊回归分析模型，将以上研究的痛点问题融入该模型，对未来人口进行预测。

三、基本假设

1、计划生育政策保持不变；

2、不考虑突发事件(如传染病暴发、战争等)和不可抗力(如地震。海啸等)对中国人口数量造成的影响；

3、将整个中国作为一个独立的人口分布系统，国际人口的迁入迁出对我国自然增长率没有影响；

4、在中短期内的死亡率和出生率保持相对稳定。

四、符号说明

y(t)	t 时刻人口总数
k	人口相对增长率
M	环境最大容纳量
$\tilde{X},\tilde{Y}$	对称三角模糊数
d_j	模糊距离
$x^{(0)}$	原始数据序列
$x^{(1)}$	累加生成序列
$z^{(1)}$	均值生成序列
$x_k(t)$	时段 t 第 k 年龄组人口数量
$b_k(t)$	生育率
$d_k(t)$	死亡率
$p_k(t)$	存活率
W_i	权值
B_i	总和生育率
$f_A(x)$	隶属函数
E	老龄化指标

五、模型的建立与求解

5.1 基于 Gompertz 模型的人口增长和预测模型的建立与求解

首先，我们建立人口增长模型。通过调查我们发现实际生活中人口增长的规律并不符合 S 型增长曲线，于是我们选择符合客观生长规律的 Gompertz 模型建立人口增长模型。其次，在建立人口预测模型方面，我们同样选择基于 Gompertz 模型建立人口预测模型。最后，将历年的实际人口数量作为观测值，拟合得到人口预测的公式，求出未来我国人口数量。

5.1.1 模型的建立

人口增长的规律符合 S 型增长曲线，但在实际生活中，由于灾难、疾病等各种客观因素的干扰，使得人口增长的规律并不是理想中的完全对称的 S 型，因此，在这里我们如果使用简单的 Logisitic 模型进行人口预测，并不能很好的描述人口增长的实际情况。所以，我们选用当前使用较多的 Gompertz 模型用以描述生物种群生长发育规律的生长曲线模型，即建立人口增长模型。

Gompertz 增长的基本规律用公式表示如下：

$$\frac{\mathrm{dy}}{\mathrm{dt}} = kyln\frac{M}{y}$$

其中，y(t)表示在 t 时刻人口的总数量；k 为人口的相对增长率，即平均出生率减去平均死亡率；M 表示环境的最大容纳量。

因为上式为可分离变量的微分方程，因此分离变量得到：

$$y = Me^{-\beta e^{-kt}}$$

其中，$\beta = \ln\frac{M}{y_0}$，y_0为中国初始时刻的人口总数。

现利用该模型对我国人口增长规律进行预测。上述模型为三个参数(M, β, k)S型增长模型，对于三个参数的拟合可以先确定其中的 M，并通过变换将其线性化，最后由最小二乘法估计出另两个参数值。

对公式（2）取对数，得到：

$$\ln y = \ln M - \beta e^{-kt}$$

$$\ln\frac{M}{y} = \beta e^{-kt}$$

$$\ln\ln\frac{M}{y} = ln\beta - kt$$

令

$$x = \ln\ln\frac{M}{y}, A = ln\beta, B = -k,$$

则最终得到：

$$x = A + Bt$$

设 n 组观测值为$(t_i, y_i), i = 1,2, \dots, n$，令

$$x_i = lnln\frac{M}{y_i}, \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

由最小二乘法得：

$$\begin{cases} \hat{B} = \dfrac{\sum_{i=1}^{n}(t_i x_i) - n\bar{t}\bar{x}}{\sum_{i=1}^{n} {t_i}^2 - n\bar{t}^2} = \dfrac{\sum_{i=1}^{n}(t_i - \bar{t})(x_i - \bar{x})}{\sum_{i=1}^{n}(t_i - \bar{t})^2} \\ \hat{A} = \bar{x} - \hat{B}\bar{t} \end{cases}$$

则$\beta = e^{\hat{A}}, k = -\hat{B},$

$$y = Me^{-\beta e^{-kt}}$$

5.1.2 模型的求解

由中国统计年鉴，我们得到 1985-2005 年我国历年的实际人口数量。将这 21 年相应的人口数据作为观测数据，则$y_0 = 10.5851, y_1 = 10.7507$，$y_2 = 10.9300, \dots, y_{20} = 12.9988, y_{21} = 13.0756$。

令 1985 年为 0 年，则$t_0 = 0, t_1 = 1, \dots, t_{20} = 20, t_{21} = 21$。

将上述观测值数据代入拟合，由调查得到 M 取 16 亿[1]，我们得到$\beta = 0.4070, k = 0.0368$，则：

$$y = 16e^{-0.4070e^{-0.0368t}}$$

上式即为中国人口增长的近似预测公式，我们利用此公式计算出中国 1985-

2005 年的人口数量。

所得数据如表 1 所示：

表 1 1985-2005 年基本数据

年份	实际人口(亿)	预测人口(亿)	误差(亿)	误差率(%)
1985	10.5851	10.6503	0.0652	0.616
1986	10.7507	10.8080	0.0574	0.5337
1987	10.9300	10.9623	0.0324	0.2961
1988	11.1026	11.1131	0.0105	0.0951
1989	11.2704	11.2604	-0.0099	-0.088
1990	11.4333	11.4043	-0.0289	-0.254
1991	11.5823	11.5446	-0.0376	-0.325
1992	11.7171	11.6816	-0.0354	-0.303
1993	11.8517	11.8151	-0.0365	-0.308
1994	11.9850	11.9453	-0.0397	-0.331
1995	12.1121	12.0721	-0.0399	-0.33
1996	12.2389	12.1956	-0.0433	-0.354
1997	12.3625	12.3158	-0.0466	-0.378
1998	12.4761	12.4328	-0.0432	-0.347
1999	12.5786	12.5466	-0.0319	-0.254
2000	12.6743	12.6573	-0.0169	-0.134
2001	12.7627	12.7650	0.0022	0.018
2002	12.8453	12.8696	0.0243	0.1892
2003	12.9227	12.9712	0.0485	0.3756
2004	12.9988	13.0699	0.0711	0.5475
2005	13.0756	13.1658	0.0902	0.6901

根据表 1 我们得到，1985-2005 年中国的实际人口和预测人口的误差值最大为 0.0652 亿，而最小误差只有 0.0022 亿，全部误差的绝对值都保持在 0.1 亿以内，即用此公式预测出的我国人口数量的结果与实际人口数量吻合度较高，最大误差率为 0.616%。因此我们认为可以利用基于 Gompertz 模型的人口增长预测模型对 2006-2050 年的中国人口数量进行直接预测。

我们对 2006-2020 年每年进行短期预测，2020-2050 年每 5 年进行预测。预测的结果如表 2 所示：

表 2 2006-2050 人口预测值

年份	预测人口(亿)
2006	13.2589
2007	13.3492
2008	13.4369
2009	13.5219
2010	13.6044
2011	13.6843
2012	13.7619
2013	13.8370
2014	13.9098
2015	13.9803

2016	14.0487
2017	14.1148
2018	14.1789
2019	14.2409
2020	14.3010
2025	14.5734
2030	14.8039
2035	14.9985
2040	15.1623
2045	15.3000
2050	15.4154

由表 2 我们可以得到，2006-2050 年的中国人口总数呈上升趋势，每年的增长人数在 0.05-0.2 亿人之间。2010 年、2020 年、2030 年、2040 年、2050 年我国的人口总数分别为 13.6044 亿人、14.3010 亿人、14.8039 亿人、15.1623 亿人、15.4154 亿人。到 2036 年左右我国人口数量达到 15 亿，2050 年的人口数量约为 15.4154 亿。

5.2 基于灰色预测模型的出生人口性别比预测模型的建立与求解

针对我国人口的痛点问题，即出生人口性别比持续升高、老龄化进程加速以及乡村人口城镇化问题，于是我们需要对我国老龄人口数、男女人口数和乡村城镇人口数进行预测。对于此类问题，我们选择基于灰色预测模型建立出生人口性别比模型。首先，我们整理得到我国 1995-2005 年市、镇、乡的出生人口性别比的数据。其次，利用灰色预测得到我国 2006-2050 年我国市、镇、乡的出生人口性别比。最后，通过残差和级比偏差检验所建模型的准确性。

5.2.1 模型的建立

为了预测我们 2006-2050 年的市、镇、乡的出生人口性别比，我们建立灰色预测模型。灰色预测模型是在前几年数据的基础上预测出未来的数据。其特点是，灰色理论建立的是生成数据模型，不是原始的数据模型，即对原始数据作累加生成（或其他方法生成）得到近似的指数规律再进行建模的方法。

我们选用 GM(1,1)模型预测是因为 GM(1,1)表示的模型是一阶微分方程，且只含一个变量的灰色模型。灰色模型的优点在于不需要很多的数据，精度高；而且运算简便，不考虑分布规律，不考虑变化趋势。

设原始数据序列：

$$x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(11))$$

由一次累加生成序列的定义

$$x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i), k = 1,2, \dots ,11$$

对数据进行累加，得到一次累加生成序列：

$$x^{(1)} = \left(x^{(1)}(1), x^{(1)}(2), .., x^{(1)}(11)\right);$$

同理我们构造$x^{(1)}$的均值生成序列，如下所示：

$$z^{(1)}(k)=\frac{1}{2}\left(x^{(1)}(k)+x^{(1)}(k-1)\right),k=1,2,\dots 11$$

$$Z^{(1)}=\left(Z^{(1)}(2),Z^{(1)}(3),\dots,Z^{(1)}(11)\right);$$

灰色微分方程模型的灰微方程以及其对应的白化方程为公式为：

$$x^{(0)}(k)+az^{(1)}(k)=b$$

$$\frac{dx^{(1)}}{dt}+ax^{(1)}(t)=b,k=2,3,\dots 11$$

再根据最小二乘法的原理，对数据进行处理得到白化方程中未知系数 a 和 b 的估计值：

为了方便对数据进行处理，我们决定将计算过程全部切换成矩阵的形式：

构造矩阵 u，Y，B 其中$u=(a,b)^T$，

$$Y=(x^{(0)}(2),x^{(0)}(3),..x^{(0)}(11))^T;$$

因此矩阵形式为：

$$B=\begin{pmatrix}-z^{(1)}(2) & 1\\ -z^{(1)}(3) & 1\\ \dots & \dots\\ -z^{(1)}(11) & 1\end{pmatrix}$$

$$Y=Bu$$

由此，目标转变为最小二乘原理的解决，即求出一个适合的矩阵 u 使得下式达到最小值：

$$J(u)=(Y-Bu)^T(Y-Bu)$$

因此最小二乘 u 的估计值为：

$$\hat{u}=(\hat{a},\hat{b})^T=(B^TB)^{-1}B^TY$$

把估计值代入求解白化方程得到公式：

$$\hat{x}^{(1)}(k+1)=\left(x^{(0)}(1)-\frac{\hat{b}}{\hat{a}}\right)e^{-\hat{a}k}+\frac{\hat{b}}{\hat{a}},k=0,1,2,\dots,11$$

当 k=1,2,...时，由上式算得的$\hat{x}^{(1)}(k+1)$是拟合值；

当$k\geq N$时，$\hat{x}^{(1)}(k+1)$为预报值。

由此我们可以得出预测的累加生成数，依据之前阐述的累加生成数的概念，采用累减的办法，我们就很容易能得到需要预测的值。

但是根据此方法算得的值不能立刻作为结论，需要通过残差检验和级比检验之后才能保证数据的可靠性，这样的预测在现实生活中也才具有实际意义。

整理 1995-2005 年我国市、镇、乡的出生人口性别比数据如表 3 所示：

表 3 1995-2005 年出生性别比

年份	市男女出生比例（女 100 计）	镇男女出生比例（女 100 计）	乡男女出生比例（女 100 计）
1995	111.92	115.62	117.75
1996	111.68	111.68	117.7

1997	108.81	125.9	118.85
1998	110.68	108.73	119.98
1999	110.27	118.4	122.03
2000	113	116.3	119.3
2001	109.28	116.02	117.59
2002	111.37	123.12	122.11
2003	112.06	110.97	120.9
2004	114.44	126.9	122.21
2005	113.92	117.21	121.21

通过上表可以看出，无论是在经济发展较快的城市还是经济相对落后的乡镇中，男女出生比例都大于 100，即说明了各地区的出生人数都是男孩较多而女孩较少。横向观察此表，我们发现，1995-2005 年每年的性别出生比的最大数值都出现在乡镇中，反映出乡镇的男女出生比例较城市不均衡，可见乡镇“重男轻女”的观念较城市严重。纵向观察表 3，我们又发现，1995-2005 年城市男女出生比的数值呈波动变化，且波动幅度较小；而乡镇的性别出生比的数值变化幅度较大，其中最大已经达到 126.9。

5.2.2 模型的求解

首先，我们对我国城市 2006 年的出生性别比进行灰色预测。

原始数列：

$$x^{(0)} = (111.92\ \ 111.68\ \ 108.81\ \ 110.68\ \ 110.27\ \ 113\ \ 109.28\ \ 111.37\ \ 112.06\ \ 114.44\ \ 113.92)$$

累加生产序列：

$$x^{(1)} = (111.92\ \ 223.6\ \ 332.41\ \ 443.09\ \ 553.36\ \ 666.36\ \ 775.64\ \ 887.01\ \ 999.07\ \ 1113.51\ \ 1227.43)$$

均值生成序列：

$$z^{(1)} = (167.76\ \ 278.01\ \ 387.75\ \ 498.23\ \ 609.86\ \ 721\ \ 831.33\ \ 943.04\ \ 1056.3\ \ 1170.5)$$

根据最小二乘法拟合得到：

$$u = (\ 0.17956609\ \ 8.0649779)$$

根据灰色预测的白化方程可以得到市场占有率的增长模型满足以下曲线方程：

$$\hat{x}^{(1)}(k+1) = \left(x^{(0)}(1) - \frac{\hat{b}}{\hat{a}}\right)e^{-\widehat{ak}} + \frac{\hat{b}}{\hat{a}} = 30377.0e^{0.003607t} - 30266.0$$

由此我们可以得到：

$$\hat{x}^{(1)}(16) = 1341.23$$

由一次累加可以得到：

$$\hat{x}^{(0)}(16) = 113.8$$

因此我们得到 2006 年我国城市出生人口性别比约为 113.8。

同理，我们可以得到 2006-2050 年城、镇、乡出生人口性别比如表 4 所示：

表 4 2006-2050 我国出生人口性别比

年份	市男女出生比例	镇男女出生比例	乡男女出生比例
2006	113.8	120.3	122.2
2007	114.2	120.8	122.5
2008	114.6	121.3	122.8
2009	115	121.8	123.2
2010	115.5	122.3	123.6
2011	115.8	122.9	124
2012	116.3	123.4	124.3
2013	116.7	123.8	124.6
2014	117.1	124.5	125.1
2015	117.5	124.9	125.4
2016	118	125.5	125.8
2017	118.3	126	126.1
2018	118.9	126.5	126.5
2019	119.2	127.1	126.9
2020	119.7	127.6	127.3
2025	121.9	130.4	131
2030	122.2	130.8	131.4
2035	122.8	131.5	131.9
2040	123.1	132	132.2
2045	123.7	132.5	132.6
2050	124	133.1	132.9

通过表 4 我们发现，2006-2050 年我国城、镇、乡的出生人口性别比都呈增长趋势，且增长幅度较平稳，没有出现大的折点，说明我国未来城、镇、乡的出生人口性别比呈缓慢增长的趋势。城市的出生人口性别比在三个地区内最小，在 2020 年左右达到 120。在 2017 年之前城镇的出生人口性别比较乡村低，在 2017-2020 年之间城镇的出生人口性别比较乡村高，在 2020 年之后两者的出生人口性别比基本持平，都在 2025 年左右达到 130。尤其值得关注的是，出生人口性别比例已经超过了 112，远远超过了正常的性别比例[4]，性别比例失衡带来了人口结构失衡等严重的社会问题。

5.2.3 模型的检验

1、残差检验

首先定义残差的计算公式如下：

$$\varepsilon(k)=\frac{x^{(0)}(k)-\hat{x}^{(0)}(k)}{x^{(0)(k)}},k=1,2,\dots 5$$

这里的$\hat{x}^{(0)}(1)=x^{(0)(1)}$，如果$\varepsilon(k)<0.1$，则可以认为达到一般要求；如果$\varepsilon(k)\geq 0.1$，则可以认为达到较高要求。

2、级比偏差值检验

级比$(\lambda(k))$和级比偏差值$(\rho(k))$的定义如下：

$$\lambda(k)=\frac{x^{(0)}(k-1)}{x^{(0)}(k)}, k=1,2,\dots 5$$

$$\rho(k)=1-\frac{1-0.5a}{1+0.5a}\lambda(k)$$

一般来说如果$\rho(k)<0.1$，则认为达到了较高的要求，我们通过计算整理，我国 1995-2005 年城市男女出生比例的偏差值如表 5 所示：

表 5 2006-2050 我国城市出生人口性别比检验

年份	残差$x^{(0)}-x(0)$	级比偏差值$\rho(k)$
1995	0	
1996	1.93094	-0.00577
1997	-1.33561	-0.03008
1998	0.13642	0.01334
1999	-0.67300	-0.00734
2000	1.65614	0.02063
2001	-2.46616	-0.03778
2002	-0.77993	0.01522
2003	-0.49514	0.00257
2004	1.47817	0.01726
2005	0.55002	-0.00819

通过对级比偏差值的分析可以得出绝对值最大的级比偏差值为 0.03778，很显然已经达到了最高的要求。所以灰色预测模型的预测值有较高的准确性。

同理，我们对 2006-2050 年我国乡镇出生人口性别比进行检验，整理数据得到表 6 和表 7：

表 6 2006-2050 我国城镇出生人口性别比检验

年份	残差$x^{(0)}-x(0)$	级比偏差值$\rho(k)$
1995	0	
1996	1.07763	1.00483
1997	-0.98734	0.6493
1998	0.44837	0.01522
1999	1.73224	-1.2239
2000	-0.44872	0.4493
2001	-1.98774	-0.00819
2002	0.44732	1.23974
2003	-0.7764	0.38872
2004	1.30042	1.47817
2005	0.8847	0.33948

表 7 2006-2050 我国乡村出生人口性别比检验

年份	残差$x^{(0)}-x(0)$	级比偏差值$\rho(k)$
1995	0	
1996	1.73224	0.39754
1997	1.22394	1.97543

1998	-0.03008	0.39224
1999	0.01334	-0.98734
2000	1.0394	0.44837
2001	0.38872	-1.39674
2002	-1.3384	0.44837
2003	0.98322	1.73224
2004	-1.29375	-0.44872
2005	-0.34872	0.49862

由此我们得出，2006 年我国市、镇、乡的出生人口性别比分别为 113.8、120.3、122.2。其中城市出生人口性别比最小，乡村出生人口性别比最大。2006-2050 年市、镇、乡的出生人口性别比都呈增长趋势，且增长较平稳，没有出现大的折点。到 2050 年我国市、镇、乡的出生人口性别比分别为 124、133.1、132.9。

3、置信区间

置信区间是指由样本统计量所构造的总体参数的估计区间,它展现的是这个参数的真实值有一定概率落在测量结果周围的程度,是被测量参数的测量值的可信程度。一般来说，置信水平越高，所对应的置信区间就会越大。

因为本题数据明显是大样本数据，在不知道总体方差的情况下，我们选择枢轴量如下公式：

$$\mathrm{t}=\frac{\bar{x}-u}{s/\sqrt{n}}\sim t(n-1)$$

所以我们可以得到移动端考研产品的价格估计区间为如下公式：

$$\left(\bar{x}-t_{\frac{\alpha}{2}}\frac{S_{n-1}}{\sqrt{n}},\bar{x}+t_{\frac{\alpha}{2}}\frac{S_{n-1}}{\sqrt{n}}\right),$$

设置信度$\alpha=0.05$，S_{n-1}为产品的样本方差，$\bar{x}$为产品的样本均值，整理数据如表 8 所示：

表 8 城市人口性别比置信区间

年份	城市人口性别比置信区间
2006	(108.4,132.8)
2007	(102.98,130.47)
2008	(104.87,128.76)
2009	(106.87,129.08)
2010	(110.97,136.63)
2011	(107.66,133.62)
2012	(103.62,134.65)
2013	(108.34,143.86)
2014	(103.4,135.65)
2015	(105.66,126.75)
2016	(114.98,138.08)
2017	(115.38,138.42)
2018	(111.07,132.76)
2019	(113.45,126.05)
2020	(112.88,130.43)
2025	(113.98,132.45)
2030	(118.64,129.04)

2035	(120.63,139.02)
2040	(118.38,136.64)
2045	(119.34,129.42)
2050	(119.82,135.03)

由此我们发现，通过灰色预测法得出的 2006-2050 年出生人口性别比的结果是比较准确的。

5.3 基于 Leslie 修正模型的城镇化预测模型的建立与求解

我国较突出的问题除了出生人口性别比持续升高之外，还包括老龄化进程加速、乡村人口城镇化等问题。在这里，我们先对我国乡村人口城镇化问题进行分析。首先，我们选择利用 Leslie 修正模型预测出中国 2006-2050 年的城镇人口数。其次，通过城镇人口数和总人口数计算城镇化率。最后，通过城镇化率分析中国未来城镇化问题的走向。

5.3.1 模型的建立

Leslie 修正模型是利用离散时间变量和离散年龄尺度以及某一初始年的人口指标数来对未来人口数目进行的预测。由此我们发现，合理的选择初始年是正确进行预测的关键[3]，为了准确预测 2006-2050 年的相关数据，我们选择 2001-2004 年的相关人口数据进行初步处理分析。

1、人口总数

通过附件所给数据，我们得到 2001-2005 年我国人口总数的观测值，其变化趋势如图 1 所示：

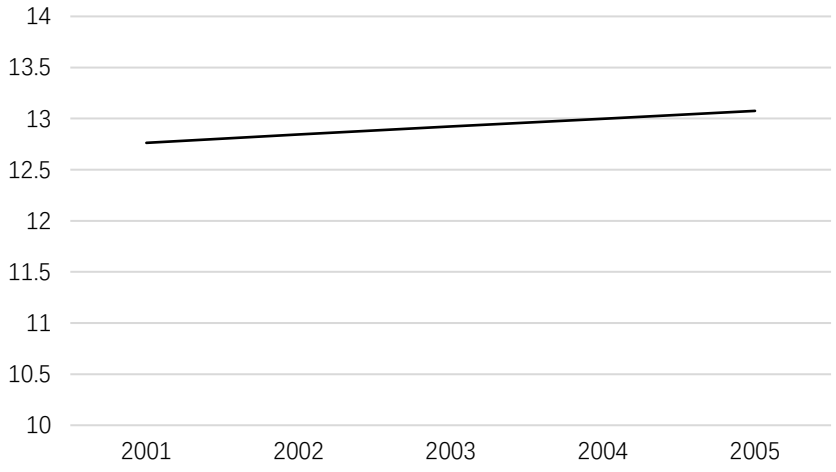

图 1 2001-2005 年我国人口总数

由图 1 我们发现，2001-2005 年我国的人口总数基本稳定，可以认为没有发生战争、灾难等异常现象。

2、总和生育率

总和生育率是指在一定时期内各年龄组妇女生育率的总和[4]，说明每名妇女按照某一年的各年龄组生育率度过育龄期，平均可能生育子女数是衡量生育水平最常用的指标之一。

整理数据如表 9 所示：

表 9 总和生育率

年份	城市总和生育率	城镇总和生育率	乡村总和生育率
2001	0.92648	1.27797	1.65371
2002	1.04831	1.34744	1.68698
2003	0.9521	1.317	1.6769
2004	0.96053	1.20339	1.65267
2005	1.00208	1.18864	1.60399

为了更清晰直观的反应 2001-2005 年各地区总和生育率的变化趋势，我们得到图 2-图 4：

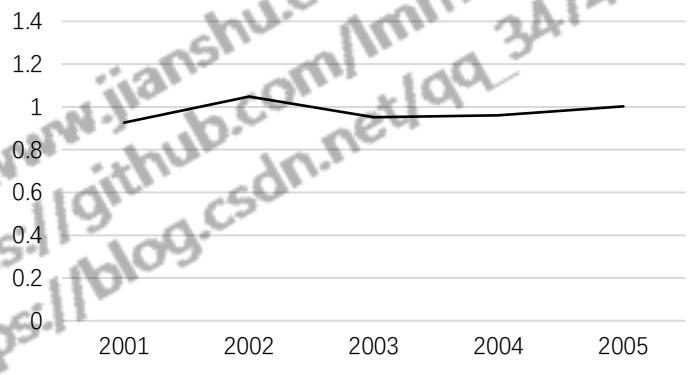

图 2 城市总和生育率

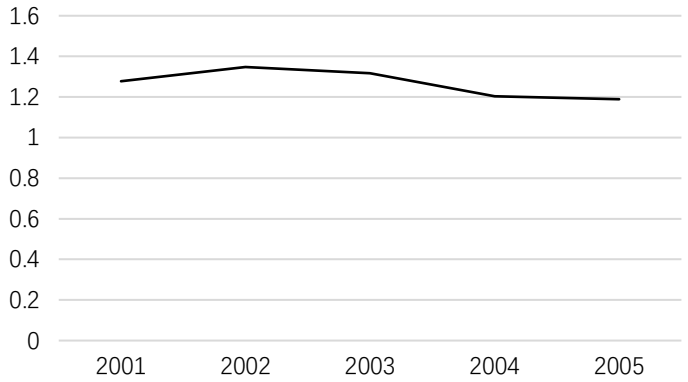

图 3 城镇总和生育率

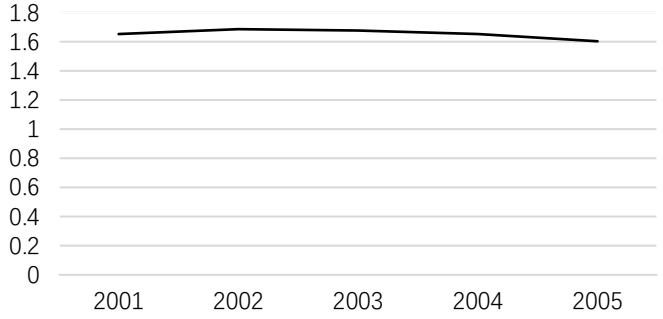

图 4 乡村总和生育率

由此我们发现，2001-2005 年城市和城镇的总和生育率较乡村总和生育率有较大的波动。城市总和生育率较低，数值在 1.0 上下波动；城镇总和生育率较高，几乎维持在 1.2-1.4 之间；乡村总和生育率最高，数值维持在 1.6 附近，且发展趋

势平缓，几乎呈稳定状态。

3、死亡率

将全国人口分为三个类别，即“城市人口”、“城镇人口”、“乡村人口”，由此可以得出某人口类别某年龄段的人口自然死亡率[5]。

$$\text{人口自然死亡率} = \text{男性比例} \times \text{男性死亡率} + \text{女性比例} \times \text{女性死亡率}$$

根据附件中所给数据，我们整理得到 2005 年各个年龄段的死亡率，如附录所示。

但是，由于我国城镇化的情况较复杂，不能单纯的使用 Leslie 模型求解，因此我们以 Leslie 模型为基础，进行适当的修正，得到 Leslie 修正模型，建立基于 Leslie 修正模型的城镇化预测模型。

1、Leslie 模型

将全国总人口数按年龄大小等间隔的分成 91 个年龄组，（这里我们记 90+为 90）间隔为一年，与年龄的离散化相对应，时间也为离散化时段，并且时段的间隔与年领区大小相等，即以一年为一个时段。人口数量是通过女性个体的生育而增长的，因此以女性人口为研究对象来预测未来女性的人口总数。

记时段 t 第 k 年龄组的人口数量$x_k(t), k = 0,1,\dots,90; t = 1,2,\dots$，时段 t 第 k 年龄组的女性生育率为$b_k(t)$，即第 k 年龄组每一妇女在一个时间段内平均生育率的数量。时段 t 第 k 年龄组的死亡率为$d_k(t)$，即第 k 年龄组在一个时间段内死亡数与总数之比，同时也可以得到存活率$p_k(t) = 1 - d_k(t)$。根据在稳定环境下假设$b_k(t)$和$d_k(t)$不随时段 t 变化的合理性，由年鉴得到$b_k(t)$和$d_k(t)$。根据 Leslie 模型的思想：时段 t+1 第 1 年龄组人口数量是时段 t 各年龄组出生人口数量之和，即：

$$x_0(t+1) = \sum_{k=0}^{90} b_k(t)x_k(t)$$

时段 t+1 第 k+1 年龄组的人口数量是由时段 t 第 k 年龄组存活的人数，即：

$$x_{k+1}(t+1) = p_k(t)x_k(t), \mathrm{k} = 0,1,\dots,90$$

其中：

$$x_j(0) = x_{j0}, j = 0,1,\dots,90$$

记时段 t 人口按年龄组分布向量：

$$\mathrm{x(t)} = [x_0(t), x_1(t), x_2(t), \dots, x_{90}(t)]^T$$

则有生育率b_k和存活率p_k构成矩阵：

$$\mathrm{L} = \begin{bmatrix} b_0(t) & b_1(t) & \cdots & b_{89}(t) & b_{90}(t) \\ p_0(t) & 0 & \cdots & 0 & 0 \\ 0 & p_1(t) & \cdots & 0 & 0 \\ 0 & 0 & p_2(t) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & p_{90}(t) & 0 \end{bmatrix}$$

把上述公式用以下递推式表示：

$$\mathrm{x(t+1)} = L\mathrm{x(t)}, t = 1,2,\dots$$

由此递推式可得到 Leslie 模型的预测公式：

$$\mathrm{x(t)} = L^t\mathrm{x(0)}$$

根据 L 矩阵和时段 t 年龄段的初始分布向量可以预测出时段 t 种群的总妇

女人数，再由性别比数据不难推算出总人口数。

2、Leslie 修正模型

由于 Leslie 模型中所用的$x_k(t)$只是时段 t 所有年龄组的妇女数目，预测结果也是女性总数，为了得到某一年的人口总数，必须先预测未来某一年的性别比例才可以得到人数，于是整个建模过程用了两次预测，两次预测均引入了误差，这样预测出的总人口数误差较大。基于此可改进 Leslie 模型，直接预测出未来人口总数。如果令$x_k(t)$为时段 t 年龄组的男女总数，另用符号$x_k{}'(t)$表示。$d_k(t)$为时段 t 某个年龄组人口自然死亡率，另用符号$d_k{}'(t)$表示。于是得到的存活率$p_k{}'(t) = 1 - d_k{}'(t)$[6]，未来统一量纲，将生育率作如下变化：

$$b_k{}'(t) = \frac{b_k(t) \times w(t)}{w(t) + m(t)}$$

其中，$w(t)$为时段 t 的女性人口总数，$m(t)$为时段 t 的男性人口总数。于是重新得到的 Leslie 矩阵变为：

$$L' = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ p_0{}'(t) & 0 & \cdots & 0 & 0 \\ 0 & p_1{}'(t) & \cdots & 0 & 0 \\ 0 & 0 & p_2{}'(t) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & p_{90}{}'(t) & 0 \end{bmatrix}$$

得到修正的 Leslie 模型人口预测公式为：

$$x'(\mathrm{t}) = L'^{t} x'(0)$$

根据某一年的初始值向量$x'(0)$可以预测未来的人口总数。

5.3.2 模型的求解

根据题目附件所给数据，我们得到城市人口对应的 Leslie 矩阵为：

$$L' = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0.9119 & 0 & \cdots & 0 & 0 \\ 0 & 0.9817 & \cdots & 0 & 0 \\ 0 & 0 & 0.9912 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0.7843 & 0 \end{bmatrix}$$

由此得到的城镇乡人数为：

表 10 城镇乡人数

年份	城	镇	乡
2006	3.254	2.9378	7.3933
2007	3.262	1.9436	7.446
2008	3.2702	1.949	7.4719
2009	3.2758	1.9522	7.5306
2010	3.2803	1.9597	7.5306
2011	3.284	1.9651	7.5651
2012	3.2866	1.9708	7.6024
2013	3.2883	1.9766	7.6412

2014	3.2886	1.9822	7.6796
2015	3.2871	1.9872	7.7155
2016	3.2842	1.9914	7.7487
2017	3.279	1.9948	7.7785
2018	3.2718	1.9971	7.8037
2019	3.2627	1.9982	7.8241
2020	3.2518	1.998	7.8402
2025	3.2842	1.9812	8.836
2030	3.3149	1.9867	8.897
2035	3.3567	1.9954	9.763
2040	3.3671	2.0026	10.142
2045	3.3812	2.0034	10.357
2050	3.391	2.0126	10.503

城镇化率为：

表 11 城镇化率

年份	城镇化率
2006	49.5701
2007	49.8990
2008	51.1002
2009	51.6712
2010	52.1024
2011	52.9981
2012	53.5672
2013	54.1482
2014	54.9724
2015	55.8625
2016	56.1436
2017	57.2439
2018	58.4324
2019	59.0018
2020	60.6135
2025	71.4142
2030	71.4142
2035	71.4142
2040	73.8208
2045	75.6662
2050	77.0763

由表 11 可见,在中国现有计划生育政策下，2035 年以后中国城镇化率将达到 70%以上,到 2050 中国城镇化率将达到 77.0765%。与顾朝林[11]等人得出的结论一致。

5.3.3 模型的检验

因为对 2005-2050 年的总人口数预测采用的是 Leslie 修正模型，因此需要对模型特性进行探讨，即对 Leslie 矩阵的稳定性进行分析。矩阵L'的数字特征对预测数据的性质有很大的关系。

矩阵L'的特征多项式为：

$$|L'-\lambda \mathrm{E}|=\begin{bmatrix} b_0'(t)-\lambda & b_1'(t) & \cdots & b_{89}'(t) & b_{90}'(t) \\ p_0'(t) & -\lambda & \cdots & 0 & 0 \\ 0 & p_1'(t) & \cdots & 0 & 0 \\ 0 & 0 & p_2'(t) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & p_{90}'(t) & -\lambda \end{bmatrix}$$

$$=p_0'(t)p_1'(t)\dots p_{89}'(t)b_{90}'(t)+p_0'(t)p_1'(t)\dots p_{88}'(t)b_{89}'(t)\lambda+p_0'(t)p_1'(t)\dots p_{87}'(t)b_{88}'(t)\lambda^2+\cdots+p_0'(t)b_1'(t)\lambda^{89}+b_0'(t)\lambda^{90}-\lambda^{91}$$

若记：

$$\mathrm{f}(\lambda)=\frac{p_0'(t)p_1'(t)\dots p_{89}'(t)b_{90}'(t)}{\lambda^{91}}+\frac{p_0'(t)p_1'(t)\dots p_{88}'(t)b_{89}'(t)}{\lambda^{90}}+\frac{p_0'(t)p_1'(t)\dots p_{87}'(t)b_{88}'(t)}{\lambda^{89}}+\cdots+\frac{p_0'(t)b_1'(t)}{\lambda^2}+\frac{b_0'(t)}{\lambda}$$

则矩阵L'的特征方程可变形为：$\mathrm{f}(\lambda)=1$。对函数$\mathrm{f}(\lambda)$进行分析可知$\mathrm{f}(\lambda)$在$\lambda>0$是连续。当$\lambda>0$时，$\mathrm{f}(\lambda)$单调减少，$\mathrm{f}(0)=0,\lim_{\lambda\to\infty}\mathrm{f}(\lambda)=0$。所以，$\mathrm{f}(\lambda)=1$有且仅有一个正实数特征值记为$\lambda_0$。根据佩龙和弗罗宾尼斯定理[8]可知$\lambda_0$是该矩阵模最大的特征值，其他特征值得模均严格小于它。L'是一个非负矩阵，它有主特征值λ_0和特征向量$\overrightarrow{v_0}$。设L'的特征值和特征向量为：

$$\lambda_0,\overrightarrow{v_0};\lambda_1,\overrightarrow{v_1};\dots;\lambda_{90},\overrightarrow{v_{90}}$$

利用矩阵的性质得：

$$\overrightarrow{x(t)}={\lambda_0}^t\left(c_0\overrightarrow{v_0}+c_1\frac{{\lambda_1}^t}{{\lambda_0}^t}\overrightarrow{v_1}+\cdots+c_{90}\frac{{\lambda_{90}}^t}{{\lambda_{90}}^t}\overrightarrow{v_{90}}\right)$$

有上式得：

当$\lambda_0<1$时，$\lim_{t\to\infty}\overrightarrow{x_k(t)}=\vec{0}$；

当$\lambda_0>1$时，$\lim_{t\to\infty}\sum_{k=0}^{\infty}x_k(t)=\infty$；

当$\lambda_0=1$时，$\lim_{t\to\infty}\overrightarrow{x_k(t)}=c_0\overrightarrow{v_0}$；

当$\lambda=1$时，L'将趋于稳定。因此，想要由预测公式$x'(\mathrm{t})={L'}^t x'(0)$得到未来的人口数趋于稳定，就必须使$\lambda_0$逼近于 1。根据 2001-2005 年的数据，记市、镇、乡三个人口类别L'矩阵的正实数特征值记为：$\lambda_1,\lambda_2,\lambda_3$。其中，$\lambda_1=0.9767,\lambda_2=0.9848,\lambda_3=0.9911$。全国人口的稳定性与总和生育率有关，因此对$\lambda_1,\lambda_2,\lambda_3$进行加权平均后来衡量全国人口的稳定性。

在权值确定之前，我们先了解更替水平的含义：更替水平是指同一批妇女生育女儿的数量恰好能代替她们本身。一旦达到生育更替水平，出生和死亡将逐渐趋于均衡，在没有国际迁入迁出的情况下，人口将最终停止生长，保持稳

定状态。一般认为，总和生育率为 2.1 即达到了生育更替水平[7]。据此取值 2.1 进行权值确定：

$$W_i = \frac{2.1}{2.1 - B_i}, i = 1,2,3$$

其中，B_i为 2004 年市、镇、乡的总和生育率，根据附件中数据得到$B_1 = 1.4083, B_2 = 1.3414, B_3 = 1.6870$。对权值$W_i$进行归一化处理后得到$W_1 = 0.3311, W_2 = 0.3335, W_3 = 0.3354$。于是 Leslie 修正模型的矩阵特征值为$\lambda = \sum W_i\lambda_i = 0.9842$。由于$\lambda < 1$，所以根据此模型预测得到的全国未来人口数将在某一年达到峰值后，逐渐趋于稳定并少有下降趋势。

5.4 基于神经网络的老龄化趋势预测模型的建立与求解

在进行完我国未来的从出生人口性别比预测和城镇化预测后，最后我们对老龄化趋势进行预测分析。为了准确的对我国未来老龄化趋势做出预测，我们建立基于神经网络的老龄化趋势预测模型。首先，我们将人口老龄化问题模糊化。其次，设计老龄化指标。最后，利用神经网络对我国未来老龄化趋势进行预测。

5.4.1 模型的建立

人口老龄化是指人口生育率降低和人均寿命延长导致的总人口数中因年经人口数量减小、年长人口数量增加而导致的老年人口比例相应增长的动态。两个含义：一是指老年人口相对增多，在总人口中所占比例不断上升的过程；二是指社会人口呈现老年状态，进入老年化社会。国际上的通常看法是，当一个国家或地区 60 岁以上老年人口占人口总数的 10%，或 65 岁以上老年人口占人口总数的 7%，即意味着这个国家或地区的人口处于老龄化社会。

所谓模糊性，通常是指对概念的定义以及语言意义的理解上的不确定性。现实生活中的许多概念含有某种程度的模糊性，很多分类问题也有定义不确定的性质，具有模糊性。根据题目中所给的信息我们发现，对人口老龄化中“老龄”的描述，常常不具有准确的答案。如使用“65 岁以上、70 岁以上、75 岁以上”等模糊词语来表述。对这种具有模糊性的对象采用模糊化的方法，即建立人口老龄化模糊集合，来描述和分析社会人口老龄化趋势和老龄化。

1、老龄化的模糊集合的建立

令 X 是一个点（对象）的空间，用 x 表示 X 的一个普通元素，于是$X = \{x\}$。X 的一个模糊集 A 通过一个隶属（特征）函数$f_A(x)$来刻画，$f_A(x)$使 X 内的每一个点与区间[0,1]内的一个实数相对应，用$f_A(x)$在点 x 的值来表示 x 在 A 内的隶属度。$f_A(x)$的值越是接近于 1，x 属于 A 内的程度就越大[8]。

运用上述模糊集合理论对人口年龄进行划分。首先设年龄 i 是大于 75 岁的人，其隶属于“老龄”这个概念的隶属度为 1；当$i \le 75$时，其隶属于“老龄”的隶属度为$i/75$[9]。上述设定可以表示为：

$$U_i = \begin{cases} \frac{i}{75} & i \le 75 \\ 1 & i > 75 \end{cases}$$

2、老龄化指标的设计

定义隶属度向量 U 为$(U_0, \dots, U_i, \dots)$，同时，为了能表示整个群体的老龄化程

度，还必须知道整个群体各个年龄的分布。假设总人口为数量 N，各个年龄的人的数量为N_i，则分布向量 D 为$(D_0, \dots, D_i, \dots)$，其中，$D_i = N_i/N$。定义 U 与 D 的内积 E 为老龄化指标，即：

$$\mathrm{E} = \mathrm{U} \times \mathrm{D} = \sum U_i D_i$$

3、神经网络的建立

BP 神经网络是模拟人类大脑处理和分析问题的方式方法来研究实际问题，从本质上说，它是一种黑箱建模工具，它能够通过“学习”来仿真真实系统里中的输入和输出之间的定量关系。它是一种具有三层或三层以上的多层神经网络，每一层都由若干个神经元组成，如图 10 所示，左层的每一个神经元与右层的每一个神经元都有连接，而上下神经元之间无连接。BP 神经网络按有导师学习方式进行训练，当一对学习模式提供给网络后，其神经元的激活值将从输入层经各隐含层向输出层传播，在输出层的各神经元输出对应于输入模式的网络相应。然后，按减少希望输出与实际输出误差的原则，从输出层经各隐含层，最后回到输入层逐层修正各连接权。由于这种修正过程是从输出到输入层逐层进行的，所以称它为“误差逆传播算法”。随着这种误差逆传播训练的不断修正，网络对输入模式响应的正确率也将不断提高。

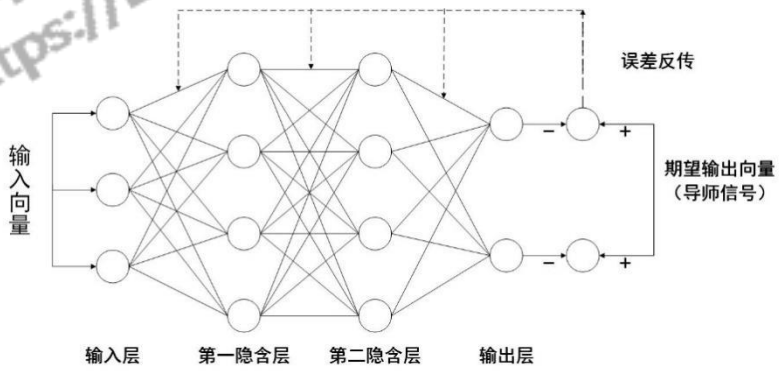

图 5 BP 网络模型结构

为了使神经网络具有某种功能，完成某项任务，必须调整层间连接权值和节点阈值，使所有样品的实际输出和期望输出之间的误差稳定在一个较小的值以内。一般地，可将 BP 网络的学习算法描述为如下步骤：

步骤一：初始化网络及学习参数；

步骤二：提供训练模式、训练网络，直到满足学习要求；

步骤三：前向传播过程：对给定训练模式输入，计算网络的输出模式，并与期望模式比较，若有误差，则执行步骤四，否则返回步骤二；

步骤四：反向传播过程：计算同一层单元的误差，修正权值和阈值，返回步骤二。

网络的学习是通过用给定的训练集训练而实现的。通常用网络的均方差误差来定量地反映学习的性能。一般地，当网络的均方差误差低于给定值时，则表明对给定训练集学习已经满足要求了。

4、基于神经网络的老龄化趋势预测模型的建立

根据上述对 BP 神经网络的描述，我们把 1995-2005 年的人口老龄化指标当作训练集，2006 年的人口老龄化指标当作测试集，再对 2007-2050 年的人口老龄化指标进行预测。

神经网络中的输入输出关系可以描述为：

$$Y_i = f(I_i)$$

$$I_i = \sum_{j=1}^{n} w_{ij} X_j - \theta_i$$

其中，X_j是神经元的输入，即是来自前级 n 个神经元的轴突的信息；θ_i是神经元的阈值；w_{ij}表示从神经元 j 到神经元 i 的连接权重；Y_i是神经元 i 的输出；$f(I_i)$是神经元的传递函数或称为激励函数，它决定神经元 i 受到输入$X_1, X_2, \ldots, X_n$的共同刺激达到阈值时以何种方式输出。

通常激励函数$f(I_i)$为非线性函数，常用的一种为 S 型函数：

$$\mathrm{Y} = \mathrm{f(I)} = \frac{1}{1 + e^{-\mu I}}$$

5.4.2 模型的求解

我们选择上述神经网络（网络结构为 1-5-1，其中隐层神经元采用正切 S 型激活函数，输出层采用 Pureline 型激活函数），采用监督学习方式，对我国人口老龄化趋势进行预测。

将人口老龄化指标函数值作为网络的输出，将时间轴作为网络的输入，构建神经网络模型。学习样本的输出为已被“模糊化”了的各年度的“老龄”隶属度向量 U 与各个年龄的人的分布向量 D 的内积即老龄化指标 E。为了防止数据产生过拟合现象[10]，我们采用三层 BP 网络建模。

利用上述老龄化指标设计方法，根据中国人口统计年鉴，可以得到有关年份我国人口老龄化的指标值，结果如表 10 所示：

表 10 我国未来老龄化指标

年份	老龄化指标
2006	0.4748
2007	0.4747
2008	0.4738
2009	0.4730
2010	0.4728
2011	0.4729
2012	0. 4730
2013	0. 4730
2014	0. 4730
2015	0. 4730
2016	0. 4730
2017	0. 4730
2018	0.4730
2019	0.4730
2020	0.4730
2025	0.4730
2030	0.4730

2035	0.4730
2040	0.4730
2045	0.4730
2050	0.4730

将表 10 中的数据整理得到图 6，如下所示：

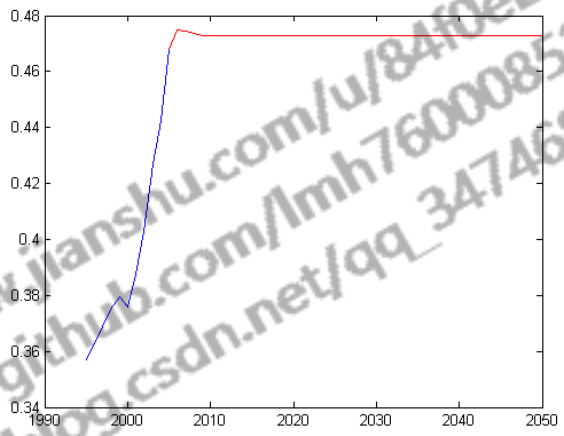

图 6 我国未来老龄化趋势

由表 10 和图 6 我们发现，未来我国人口老龄化指标的变化趋势在 2010 前之前有较大幅度的增长，2010 年之后发展较平稳，逐步稳定在 0.4730。

5.5 基于模糊线性回归分析的人口预测模型的建立与求解

根据我国 2000 年人口普查表明，我国人口统计的数量质量呈下降趋势[11]，人口漏报达到 1.81%。人口统计的数据质量一般包括总量数据的质量和结构数据的质量。总量数据的质量和结构数据的质量之间相互影响。比如某一年龄段的人口数据出现漏报或错报，可能影响总量数据的质量、性别和年龄数据的错报等；总量数据的失真则必然影响到结构数据的质量[12]。因此，我们优化之前建立的基于 Gompertz 模型的人口增长预测模型，解决带有模糊信息的动态预测问题，建立基于模糊线性回归分析的人口预测模型。

5.5.1 模型的建立

1、模糊回归分析模型

为了解决带有模糊信息的动态预测问题，我们使用模糊回归预测技术建立基于模糊线性回归分析的人口预测模型。模糊线性回归模型描述的是模糊变量和模糊自变量之间的线性相关程度。在模糊线性回归中，其自变量或因变量或二者都不是一个确切的实数，而是一个区域或一个模糊数。与线性回归类似，我们假设因变量是若干自变量的线性组合，在我国人口数量预测中，我们选择设立模糊回归模型：

$$\tilde{Y} = A_0 + A_1\widetilde{X_1} + A_2\widetilde{X_2} + A_3\widetilde{X_3} + A_4\widetilde{X_4}$$

其中，$\tilde{Y}$、$\tilde{X}$均为对称三角模糊数，可以分别表示为$\tilde{Y} = (y, u)$，y 和 u 分别是对称三角模糊数$\tilde{Y}$的中心和广度；$\tilde{X} = (x, v)$，x 和 v 分别是对称三角模糊数$\tilde{X}$的中心和广度。

与统计中的最小二乘回归分析类似，我们用模糊数代替随机变量，找出变量之间的不确定关系。$\tilde{Y}$、$\tilde{X}$的隶属函数分别表示为：

$$\tilde{Y}(a)=\begin{cases}1-\dfrac{y-a}{u} & y-u\le a\le y\\ 1+\dfrac{y-a}{u} & y\le a\le y+u\\ 0 & \text{其他}\end{cases}$$

$$\tilde{X}(b)=\begin{cases}1-\dfrac{x-b}{v} & x-v\le b\le x\\ 1+\dfrac{x-b}{v} & x\le b\le x+v\\ 0 & \text{其他}\end{cases}$$

又因为$\tilde{Y}$、$\tilde{X}$为对称三角模糊数，所以该模型的确定在于参数 y、u 和 x、v 的确定。

2、模糊最小二乘法(FLS 法)

FLS 法的基本思想是通过回归系数的选取，使得所有模糊因变量与因变量观测值之间的模糊距离最小，一般使用平方和计算。FLS 法的拟合原则如下：

给定观测值$\widetilde{Y_j}=(y_i,u_i)$与拟合值$\widetilde{Y_j}=\left(\sum_{j=1}^{n}(A_0+A_1x_j),\sum_{j=1}^{n}(A_1v_j)\right)$，

$\mathrm{j}=1,2,\ldots,\mathrm{n}$之间的模糊距离为：

$$d_j\triangleq d_j\left(\widetilde{Y_j},\widetilde{Y_j}\right)=\sqrt{\sum_{j=1}^{n}(y_j-(A_0+A_1x_j))^2+\sum_{j=1}^{n}(u_{\mathrm{j}}+A_1v_j)^2}$$

从而可以定义模糊距离平方和的距离为：

$$d_2\triangleq\sum_{j=1}^{n}{d_{\mathrm{j}}}^2=\sum_{j=1}^{n}(y_j-(A_0+A_1x_j))^2+\sum_{j=1}^{n}(u_{\mathrm{j}}+A_1v_j)^2$$

通过求解公式的最小值，即可以得到该模型中回归系数的 FLS 估计。

5.5.2 模型的求解

首先，对自变量(年份)和因变量(人口数)数据进行模糊化处理。进行模糊化处理的原因为：

(1)人口的统计不一定是在年末进行的，无论从时间上还是统计数量上，都应该是一个模糊数；

(2)由于各种原因，现实中人口的瞒报、错报和漏报现象比较严重，有必要对人口数量进行模糊化处理。

通过查阅资料，我们得到 1995-2005 年的人口总数如表 11 所示：

表 11 1995-2005 年人口数量

年份	人口数量(亿)
1995	12.1121

1996	12.2389
1997	12.3625
1998	12.4761
1999	12.5786
2000	12.6743
2001	12.7627
2002	12.8453
2003	12.9227
2004	12.9988
2005	13.0756

其次，我们对人口数据观测值进行模糊三角化处理。为了体现年份之间的关系，被解释的变量取作(Y-1994)。

处理后的数据如表 12 所示：

表 12 模糊三角化处理结果

年份	$\tilde{X}=(x,v)$		$\tilde{Y}=(y,u)$	
	X-1994	v	Y	u
1995	1	1/2	121121	1407
1996	2	1	122389	1236
1997	3	1	123626	893
1998	4	1	124761	965
1999	5	1	125786	936
2000	6	1	126743	783
2001	7	1	127627	1174
2002	8	1	128453	976
2003	9	1	129227	763
2004	10	1	129988	1211
2005	11	1/2	130756	730

根据表 11 和表 12 的数据，利用 FLS 方法得到模糊回归模型(其中自变量取年份的序列号(年份-1994))

$$\tilde{Y}=69944.63+5003.66\widetilde{X_1}+100.3356\widetilde{X_2}+11285.74\widetilde{X_3}+10725.23\widetilde{X_4}$$

由上式得到模糊样本观测数据中心和广度的回归模型分别为：

$$\begin{cases} \mathrm{Y}=69944.63+5003.66X_1+100.3356X_2+11285.74X_3+10725.23X_4 \\ \mathrm{U}=60036.3\mathrm{V} \end{cases}$$

根据上式可以看出，人口总量的模糊三角数的中心值是年份模糊三角数的中心值的线性函数，人口总量的模糊三角数的广度是年份模糊三角数的广度的线性函数。

根据计算得到 2006-2050 年中国的人口总数，我们经过整理，结果如图 7 所示：

表 14　2006 至 2050 年人口预测数量

年份	预测人口(亿)
2006	13.2589
2007	13.3492
2008	13.4369
2009	13.5219

2010	13.6044
2011	13.6843
2012	13.7619
2013	13.8370
2014	13.9098
2015	13.9803
2016	14.0487
2017	14.1148
2018	14.1789
2019	14.2409
2020	14.3010
2025	14.5734
2030	14.8039
2035	14.9985
2040	15.1623
2045	15.3000
2050	15.4154

由表 14 我们可以得到，2006-2050 年的中国人口总数呈上升趋势，每年的增长人数在 0.05-0.2 亿人之间。2010 年、2020 年、2030 年、2040 年、2050 年我国的人口总数分别为 13.6044 亿人、14.3010 亿人、14.8039 亿人、15.1623 亿人、15.4154 亿人。到 2036 年左右我国人口数量达到 15 亿，2050 年的人口数量约为 15.4154 亿。相比于模型一，此数据更加精确。

六、模型的评价与推广

6.1 模型的优点与缺点

本文在建立基于 Gompertz 模型的人口增长模型时，考虑到由于客观因素人口增长规律并不符合 S 型曲线的事实，我们决定采用 Gompertz 模型而不是 Logisitic 模型对人口未来增长趋势进行预测，有较强的拟合度和可信度。针对不同年龄段的城镇化预测问题上，考虑了不同群体的生育水平和死亡率，建立了基于 Leslie 模型的城镇化预测模型，有较高的精度。基于人口增长规律由于客观因素并不符合 S 型曲线的事实，我们决定采用 Gompertz 模型而不是 Logisitic 模型对人口未来增长趋势进行预测，该模型也是当前使用较多的描述生物种群生长发育规律的曲线模型，有较强的拟合度和可信度。基于模糊回归分析的人口预测模型不仅考虑了人口普查中的错报漏报情况，也将我国三个重要的痛点问题融入该模型，使得结果较准确。

然而 Leslie 模型只适用于中短期人口数量的预测，进行长期预测则误差较大。在假设中，我们假设死亡率基本不变，但但国民素质、医疗水平以及国家对食品安全重视度的提高等因素会引起人口死亡率下降 。中短期死亡率的误差对总人口预测影响不大，但在长期预测时，由于递推公式的作用、随着使用死亡率数据的迭代次数增加，人口结构预测的偏差就会越大。另外，我们假设国际迁入迁出对我国人口自然增长率没有影响，所以该模型是在较理想的环境下进行讨论的，无法从更深层次进行研究。

6.2 模型的推广

在基于 Gompertz 模型的人口增长预测模型中，Gompertz 模型在医学、软件开发、交通运输等领域的应用都非常广泛。基于模糊线性回归的人口预测模型，不仅能利用动态模糊很好的解决人口预测问题，还能很好的进行价格预测，比如机票价格、股票价格等等。

参考文献

[1]国家统计局.中国统计年鉴[Z].北京：中国统计出版社，2007

[2]国家统计局人口和社会科技统计司.中国人口统计年鉴[M].北京:中国统计出版社,2007

[3]任强,侯大道.人口预测的随机方法：基于 Leslie 矩阵和 ARMA 模型[J].人口研究,2015,35(2):41-58

[4]王焕清.不同计划生育政策下的我国人口预测研究[J].统计与决策,2016(5)

[5]刘贵文,杨建伟,邓恂.影响中国城市化进程的经济因素分析[J].城市发展研究,2016,13(5):9-12

[6]Suyeon Kim.Supplement 6 The Perron-Forbenius Theorem[J]. Wellesley-Cambridge Press.2015

[7]华逸群,曹健.机票价格预测的模糊时间序列方法[J].上海：上海交通大学计算机科学与工程系,2016

[8]Huarng K.Heuristic models of fuzzy time series for forecasting[J].Fuzzy Sets and Systems,2015,123(3):369-386

[9]陈卫.国际视野下的中国人口老龄化[J].北京:中国人民大学人口与发展研究中心,2017

[10]Alho J.Migration,Fertility,and Aging in Stable Populations[J]. Demography,2017,45(3):641-650

[11]张为民.对我国人口统计数据质量的几点认识[J].人口研究,2016(9)

[12]顾朝林,管卫华,刘合林. 中国城镇化 2050:SD 模型与过程模拟[J]. 中国科学:地球科学,2017,47(07):818-832. [2017-08-09].

附录

附录一：

```
function
x2=logistic(A,t)
n=length(A);
B=ones(n-1,1);
for i=2:n
B(i-1)=A(i)-A(i-1);
end
x=A(2:n,:);
y=B./x;
xx=x-mean(x);
 yy=y-mean(y);
sxy=sum(xx.*yy);
sxx=sum(xx.^2);
b1=sxy/sxx;
b0=mean(y)-b1*mean(x);
xm=-b0/b1;
x0=A(1);
r=b0;
t=t-1980; f
or i=1:t
x1(i+1)=xm/(1+(xm/x0-1)*exp(-r*i));
end
x1(1)=x0;
x1=x1';
y1=[1980:1:t+1980];
y1=y1';
x2=[y1,x1];
```

附录二：

```
function B=jianyan(A)
n=length(A);
 B=zeros(1,n-1);
for i=2:n
    B(i-1)=A(i-1)/A(i);
if B(i-1)>exp(-(2/(n+1)))&B(i-1)<exp(2/(n+1))
B(i-1)=1;
else
B(i-1)=0;
end
end
```

附录三：

```
function
y=yuce(A,t)
n=length(A);
D=ones(n,1);
C=ones(n-1,1);
for i=1:n
D(i)=sum(A(1:i));
end
for i=2:n
C(i-1)=0.5*D(i)+0.5*D(i-1);
end C=-C;
E=ones(n-1,1);
B=[C,E];
B=vpa(B);
a=inv(B'*B);
b=B';
c=a*b;
Y=A(2:n,:);
d=c*Y;
a=d(1);
b=d(2);
c=A(1);
t=t-1980;
for i=1:t
x1=(c-b/a)*exp(-a*i)+b/a;
x2=(c-b/a)*exp(-a*(i-1))+b/a;
x(i+1)=x1-x2;
end
x(1)=(c-b/a)*exp(-a*0)+b/a; a=0;
for i=1:t
for j=i
a=a+x(i);
end
y(i+1)= 9.8706+a; end
y(1)= 9.8706;
```

附录四：

```
x=[
0.3568
0.3632
0.3687
0.3748
0.3797
```

```
0.3758
0.3877
0.4029
0.4269
0.4425
0.4681
]';
lag=3;
iinput=x;
n=length(iinput);
inputs=zeros(lag,n-lag);
for i=1:n-lag
    inputs(:,i)=iinput(i:i+lag-1)';
end
targets=x(lag+1:end);
hiddenLayerSize = 4;
net = fitnet(hiddenLayerSize);
net.divideParam.trainRatio = 9/11;
net.divideParam.valRatio = 1/11;
net.divideParam.testRatio = 1/11;
 [net,tr] = train(net,inputs,targets);
yn=net(inputs);
errors=targets-yn;
figure, ploterrcorr(errors)
figure, parcorr(errors)
figure,plotresponse(con2seq(targets),con2seq(yn))
figure, ploterrhist(errors)
figure, plotperform(tr)
fn=45;
f_in=iinput(n-lag+1:end)';
f_out=zeros(1,fn);
for i=1:fn
    f_out(i)=net(f_in)
    f_in=[f_in(2:end);f_out(i)];
end
figure,plot(1995:2005,iinput,'b',2005:2050,[iinput(end),f_out],'r')
```