Operations Research and Fuzziology 运筹与模糊学, 2022, 12(3), 1081-1092
Published Online August 2022 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2022.123114

Hans汉斯

基于灰色－马尔可夫模型对中国结婚人数的预测

张　娅

贵州大学数学与统计学院，贵州　贵阳

收稿日期：2022年7月19日；录用日期：2022年8月17日；发布日期：2022年8月24日

摘　要

中国的老龄化越来越严重，而中国的结婚人数又持续走低，这对中国的年龄结构有很大的影响，劳动力市场也会不完善。因此预测中国的结婚人数有利于制定相关的政策以及采取相应的措施去应对当前形势。本文主要采用灰色－马尔可夫链进行预测，并将其与直接由GM(1,1)模型得到的预测值以及时间序列预测所得的结果进行对比。结果表明利用灰色－马尔可夫链得到的预测值比直接用GM(1,1)模型以及时间序列预测得到的值要好。

关键词

灰色－马尔可夫链，时间序列，预测，结婚人口数

Prediction of the Number of Marriages in China Based on Gray-Markov Model

Ya Zhang

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Jul. 19th, 2022; accepted: Aug. 17th, 2022; published: Aug. 24th, 2022

Abstract

China's aging is getting more and more serious, and the number of marriages in China continues to decline, which has a great impact on China's age structure and the labor market will be imperfect. Therefore, predicting the number of marriages in China is conducive to making relevant policies and taking corresponding measures for the current situation. In this paper, gray-Markov chain is mainly used for forecasting, and the results are compared with those obtained from time series

文章引用: 张娅. 基于灰色-马尔可夫模型对中国结婚人数的预测[J]. 运筹与模糊学, 2022, 12(3): 1081-1092.
DOI: 10.12677/orf.2022.123114

forecasting. The results show that the predicted values obtained by using gray-Markov chain are better than those obtained by using GM(1,1) model and time series directly.

Keywords

Gray-Markov Chain, Time Series, Prediction, Married Population

1. 引言

中国的结婚人数持续走低，更是在 2021 年创下了历史新低。文献[1]研究了 17~34 岁年龄段青年的婚恋观；文献[2]调查了当代女大学生的婚恋观；文献[3]研究了农村 70 后、80 后、90 后的婚恋观状况及其影响因素；文献[4]研究了新生代农民工婚恋观现状及其影响因素。这些文献都是建立在影响因素上去探究的，没有从数量上去探究未来结婚人数的变化。

文献[5]利用 BP 神经网络，灰色预测模型，回归分析法对石油的消费情况进行了预测。文献[6]在原有的灰色－马尔可夫模型的基础上提出了新的状态的划分方式。文献[7]利用多元线性回归，灰色马尔可夫模型以及指数平滑法对对石油进行了预测。

中国的结婚人口不是单纯递增递减的形式，由于原始数据集年度时间数据集，所以排除利用回归模型对数据进行预测。由于当前的未来的结婚数据只会与当前是否结婚与离婚有关，并且从数据的趋势看，中国的结婚人口数据波动比较大。灰色－马尔可夫模型可以克服较大的随机性，并且只需要少量数据就可以对未来进行预测，因此选择利用灰色－马尔可夫模型进行预测。同时，由于数据是时间序列数据，因此也选用时间序列的方法对数据进行预测，比较两者的效果。

本文是从以往数据出发来预测未来结婚人数的高低。这有利于国家制定相应的政策来缓解当前结婚率低的压力。结婚人数创新低有很多主观和客观原因。主观原因是 21 世纪人们的受教育程度普遍提高，择偶标准发生变化。不再是古代的父母之命，媒妁之言。客观原因是社交圈狭小，没有机会去认识新的人。不婚主义者是少数，人们普遍不婚主要是受到物质和社交圈的限制而导致不婚。因此本文的研究具有一定的现实意义。

2. 模型介绍

2.1. GM(1,1)

第一：将所有的原始数据设置为与时间相关的序列

$$X^{(0)}=\left(x^{(0)}(1),x^{(0)}(2),\cdots,x^{(0)}(n)\right) \tag{1}$$

第二：对原始序列进行累加随时间递增的单调序列，让 $X^{(1)}$ 作为 $X^{(0)}$ 的一个累加序列

$$X^{(1)}=\left(x^{(1)}(1),x^{(1)}(2),\cdots,x^{(1)}(n)\right) \tag{2}$$

$$X^{(1)}(k)=\sum_{i=0}^{k}X^{(0)}(i) \tag{3}$$

第三：令 $Z^{(1)}$ 是由 $X^{(1)}$ 生成的

$$Z^{(1)}=\left(Z^{(1)}(1),Z^{(1)}(2),\cdots,Z^{(1)}(n)\right) \tag{4}$$

$$Z^{(1)}(k)=0.5X^{(1)}(k)+X^{(1)}(k-1) \tag{5}$$

第四：确定 GM(1,1)的精确方程，即 GM(1,1)的灰色微分方程模型为：

$$x^{(1)}(k)+aZ^{(1)}(k)=b \tag{6}$$

其中 a 称为发展系数，b 为灰色作用量，假设 $\hat{a}$ 是要估计的系数，并且

$$\hat{a}=(a,b)^{\mathrm{T}},\ \hat{a}=\left(B^{\mathrm{T}}B\right)^{-1}B^{\mathrm{T}}Y_n \tag{7}$$

其中

$$B=\begin{pmatrix}-z^{(1)}(2) & 1\\ -z^{(1)}(3) & 1\\ \vdots & \vdots\\ -z^{(1)}(n) & 1\end{pmatrix},\ y_{(n)}=\begin{pmatrix}x^{(0)}(2)\\ x^{(0)}(3)\\ \vdots\\ x^{(0)}(n)\end{pmatrix}$$

$\frac{\mathrm{d}x^{(1)}}{\mathrm{d}t}+ax^{(1)}=b$，叫白化方程，也叫影子方程。

第五：根据以上步骤得到最终的预测方程

1)

$$\widehat{x^{(1)}}(k)=\left(x^{(1)}(0)-\frac{b}{a}\right)\mathrm{e}^{-ak}+\frac{b}{a} \tag{8}$$

2)

$$\widehat{x^{(1)}}(k)=\left(x^{(1)}(0)-\frac{b}{a}\right)\mathrm{e}^{-ak}+\frac{b}{a},\ k=1,2,\cdots,n \tag{9}$$

3)

$$\widehat{x^{(1)}}(k)=\left(x^{(1)}(1)-\frac{b}{a}\right)\mathrm{e}^{-ak}+\frac{b}{a},\ k=1,2,\cdots,n \tag{10}$$

4)

$$\widehat{x^{(0)}}(k+1)=\widehat{x^{(1)}}(k+1)-\widehat{x^{(1)}}(k) \tag{11}$$

2.2. 灰色－马尔可夫模型

GM(1,1)预测模型在灰色系统理论里占有重要地位，该模型研究的出发点是在自身时间序列中探索有价值的信息，发掘研究内容的规律，不需要考虑研究内容带来的影响。灰色系统模型以少量数据信息为研究对象，该模型对少量信息建模有较好的预测精度[6]。马尔可夫链是指未来的状态只与当前的状态相关联，而与过去的状态无关。马尔可夫预测模型的原理是基于当前时间的状态和状态间的转移概率来预测未来时间的状态。马尔可夫链预测与灰色系统模型不同，它弥补了灰色预测模型的不足，可以对波动性大的数据进行研究[6]。

在建立马尔可夫模型之前，需要对数据进行检验分别为级比检验和光滑度检验，级比的计算公式为：

$$\rho(k)=\frac{X^{(0)}(k-1)}{X^{(0)}(k)}(k=2,3,\cdots,n) \tag{12}$$

如果所有级比都在$\left(e^{-\frac{2}{n+1}},e^{\frac{2}{n+1}}\right)$范围之内，则可以对原始数据直接建立灰色模型。光滑比的计算公式：

$$\delta(k)=\frac{X^{(0)}(k)}{X^{(1)}(k-1)}(k=2,3,\cdots,n) \tag{13}$$

其中$X^{(1)}(k-1)=\sum_{i=1}^{k-1}X^{(0)}(i)(k=2,3,\cdots,n)$，当原始数据序列满足$\delta(k)$是$k$的递减函数时，才可以建模预测，只有当原始数据序列通过级比检验和光滑度检验时，才可以进行建模预测，否则需要对原始数据序列进行预处理。同时还要对发展系数计算发展系数$-a$，其关系见表 1：

Table 1. Application range of development coefficient $-a$ and GM(1,1) model
表 1. 发展系数$-a$与 GM(1,1)模型的适用范围

级别	$-a$	GM(1,1)模型的适用范围
1	$-a<0.3$	适用于中长期
2	$0.3<-a\leq 0.5$	适用于短期
3	$0.5<-a\leq 0.8$	适用于短期，要谨慎
4	$0.8<-a\leq 1$	残差修正
5	$-a>1$	不宜使用 GM(1,1)

马尔科夫链预测的步骤如下：

第一、计算γ值

γ值为实际值比上预测值，计算公式为：

$$\gamma=\frac{X^{(0)}(k)}{X^{(1)}(k)} \tag{14}$$

根据γ值进行状态的划分，把γ值划分为m个区间，即m个状态，任一状态区间可以表示为

$$E_i=(\gamma_i,\gamma_j)$$

γ的所有状态区间集合可以表示为$E=(E_1,E_2,\cdots,E_m)$。

第二、计算状态转移概率

$$P_{ij}(s)=\frac{M_{ij}(s)}{M_i} \tag{15}$$

第三、计算状态转移概率矩阵

$$P(s)=\begin{bmatrix}p_{11}(s) & p_{12}(s) & \cdots & p_{1m}(s)\\ p_{21}(s) & p_{22}(s) & \cdots & p_{2m}(s)\\ \vdots & \vdots & \ddots & \vdots\\ p_{m1}(s) & p_{m2}(s) & \cdots & p_{mm}(s)\end{bmatrix}$$

其中$P_{i1}(1)+P_{i2}(s)+\cdots+P_{im}(s)=1$。

第四、确定对象转移状态

马尔科夫链的下一步状态和预测对象过去的状态没有关系，只和当前的状态有关。当预测对象的状态在 E_i 时，只需观察状态转移概率矩阵中最大概率转入下一状态即可。如果有多个状态的转移概率相同，则可以具体分析应该转入哪一个状态即可。

第五、确定修正后的预测值

根据文献[6]提出的状态划分方式，每个状态区间的对应的数据不少于 7 个，且进行划分时，将两端的数据各划分为一个状态中间的数据平均分配。因此修正后的预测值的计算公式为：

$$\hat{Y}(k)=\widehat{X^{(0)}}(k)*\left(\frac{\gamma_i+\gamma_j}{2}\right) \tag{16}$$

3. 实证

本次的数据来源于中华人民共和国民政部。在建立 GM(1,1)模型之前，已对数据进行级比检验和光滑度检验，且检验都通过，可以对数据进行灰色－马尔可夫链模型进行预测。同时计算出发展系数$-a$ = 0.004164524，即原始数据可以进行中长期预测。表 2 给出了根据 GM(1,1)模型得出的相关数据，并根据 γ 值进行了状态的划分，结果见表 2：

Table 2. Simulation data of GM(1,1) model
表 2. GM(1,1)模型的模拟数据

年份	原始数据 (单位：万对)	模拟数据 (单位：万对)	原始/模拟 γ	相对误差	绝对误差	状态
1986	882.3	882.3	1	0	0	M2
1987	924.7	889.3281	1.039773735	−0.039773735	−35.3719	M3
1988	899.2	893.0395	1.006898351	−0.006898351	−6.1605	M3
1989	934.8	896.7663	1.042412053	−0.042412053	−38.0337	M4
1990	951.1	900.5087	1.056180801	−0.056180801	−50.5913	M4
1991	953.6	904.2667	1.054556139	−0.054556139	−49.3333	M4
1992	954.5	908.0404	1.051164684	−0.051164684	−46.4596	M4
1993	912.1	911.8299	1.000296218	−0.000296218	−0.2701	M2
1994	929	915.6351	1.014596317	−0.014596317	−13.3649	M3
1995	929.7	919.4562	1.011141151	−0.011141151	−10.2438	M3
1996	934	923.2933	1.011596207	−0.011596207	−10.7067	M3
1997	909	927.1464	0.980427687	0.019572313	18.1464	M2
1998	891.8	931.0156	0.957878686	0.042121314	39.2156	M2
1999	885.3	934.9009	0.946945286	0.053054714	49.6009	M2
2000	848.5	938.8025	0.903810972	0.096189028	90.3025	M1
2001	805	942.7202	0.853911903	0.146088097	137.7202	M1

Continued

2002	786	946.6545	0.830292361	0.169707639	160.6545	M1
2003	811.4	950.605	0.853561679	0.146438321	139.205	M1
2004	867.2	954.5721	0.908469879	0.091530121	87.3721	M1
2005	823.1	958.5557	0.858687711	0.141312289	135.4557	M1
2006	945	962.556	0.981761061	0.018238939	17.556	M2
2007	991.4	966.5729	1.025685698	−0.025685698	−24.8271	M3
2008	1098.3	970.6066	1.131560408	−0.131560408	−127.6934	M4
2009	1212.2	974.6572	1.24371933	−0.24371933	−237.5428	M5
2010	1241	978.7246	1.267976712	−0.267976712	−262.2754	M5
2011	1302.4	982.809	1.32518119	−0.32518119	−319.591	M5
2012	1323.6	986.9105	1.341155049	−0.341155049	−336.6895	M5
2013	1346.9	991.0291	1.359092281	−0.359092281	−355.8709	M5
2014	1306.7	995.1649	1.313048722	−0.313048722	−311.5351	M5
2015	1224.7	999.3179	1.225535938	−0.225535938	−225.3821	M5
2016	1142.8	1003.488	1.138827769	−0.138827769	−139.312	M4
2017	1063.1	1007.676	1.055001806	−0.055001806	−55.424	M4
2018	1013.9	1011.881	1.001995294	−0.001995294	−2.019	M3
2019	927.3	1016.104	0.912603434	0.087396566	88.804	M2
2020	814.3	1020.344	0.798064182	0.201935818	206.044	M1
2021	763.6	1024.603	0.745264263	0.254735737	261.003	M1

对数据进行状态的划分是灰色－马尔可夫的重点，一般用残差的相对值进行状态的划分，而本篇文章根据γ进行状态的划分，见文献[6]，由以上数据我们可把数据分为5个状态。状态区间分别为：

$$M_1:[0.75,0.92],\ M_2:[0.92,1.007],\ M_3:[1.007,1.05],\ M_4:[1.05,1.14],\ M_5:[1.14,1.36]$$

一步状态转移概率矩阵为

$$P=\begin{bmatrix} \frac{6}{7} & \frac{1}{7} & 0 & 0 & 0 \\ \frac{2}{7} & \frac{2}{7} & \frac{3}{7} & 0 & 0 \\ 0 & \frac{2}{7} & \frac{3}{7} & \frac{2}{7} & 0 \\ 0 & \frac{1}{7} & \frac{1}{7} & \frac{4}{7} & \frac{1}{7} \\ 0 & 0 & 0 & \frac{1}{7} & \frac{6}{7} \end{bmatrix}$$

利用马尔可夫模型算出预测值以后，同时以 1986~2016 年为训练集，2017~2021 作为测试集，根据时间序列中的移动平均法和两参数法，利用 R 语言对数据进行预测，相关代码见附录，代码均来源于文献[8]。不同模型的预测值见表 3；GM(1,1)、灰色－马尔可夫模型与原始数据的对比见图 1；移动平均法的预测结果见图 2；两参数模型的预测结果见图 3；移动平均法的趋势拟合见图 4；两参数模型的趋势拟合图见图 5。

Table 3. Prediction data of related models
表 3. 相关模型的预测数据

年份	原始结婚人数(万对)	GM 预测数	马尔可夫链预测值	移动平均法	两参数预测
1986	882.3	882.3	882.3		
1987	924.7	889.3281	900.078345		
1988	899.2	893.0395	907.248		
1989	934.8	896.7663	911.034245925		
1990	951.1	900.5087	977.475267		
1991	953.6	904.2667	981.554483		
1992	954.5	908.0404	985.36		
1993	912.1	911.8299	989.72		
1994	929	915.6351	930.203272485		
1995	929.7	919.4562	933.43725		
1996	934	923.2933	937.51785		
1997	909	927.1464	941.897659995		
1998	891.8	931.0156	945.82839996		
1999	885.3	934.9009	949.77556434		
2000	848.5	938.8025	952.82		
2001	805	942.7202	779.2		
2002	786	946.6545	782.457766		
2003	811.4	950.605	785.723235		
2004	867.2	954.5721	789		
2005	823.1	958.5557	792.294843		
2006	945	962.556	795.6		
2007	991.4	966.5729	981.9515034		
2008	1098.3	970.6066	986		

Continued

2009	1212.2	974.6572	1057.961194		
2010	1241	978.7246	1218.3215		
2011	1302.4	982.809	1223.40575		
2012	1323.6	986.9105	1228.638125		
2013	1346.9	991.0291	1233.638125		
2014	1306.7	995.1649	1238.786375		
2015	1224.7	999.3179	1243.956125		
2016	1142.8	1003.488	1249.147375		
2017	1063.1	1007.676	1093.8	1237	1055.7693
2018	1013.9	1011.881	1098.36684	1205.1	958.7965
2019	927.3	1016.104	1032.27	1173.185	861.8237
2020	814.3	1020.344	1036.5784956	1141.265	764.851
2021	763.6	1024.603	846.88552	1109.346	667.8782
2022			896.527625	1077.427	570.9

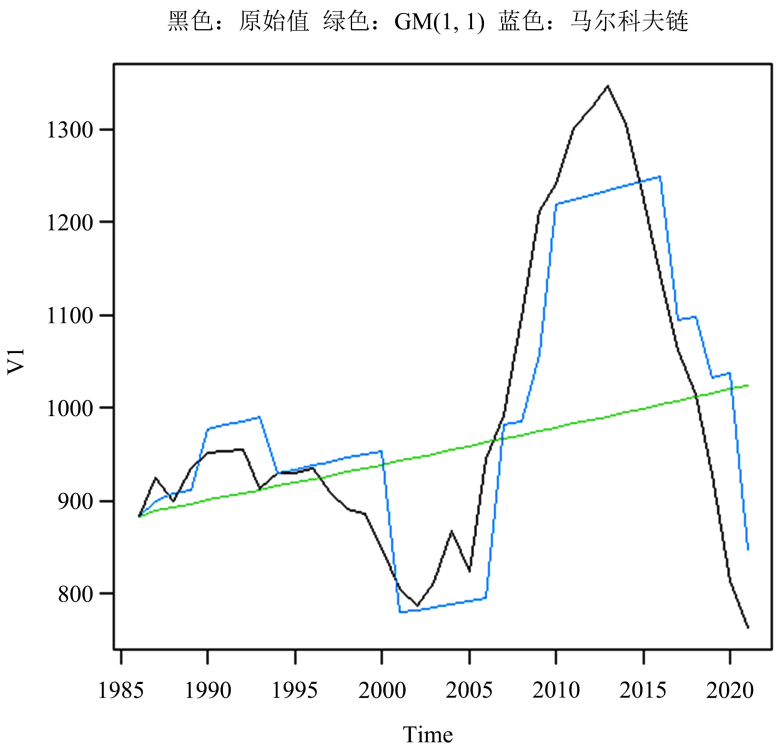

Figure 1. Comparison of raw data with GM(1,1) and Markov model predictions
图 1. 原始数据与 GM(1,1)和马尔可夫模型预测值的对比图

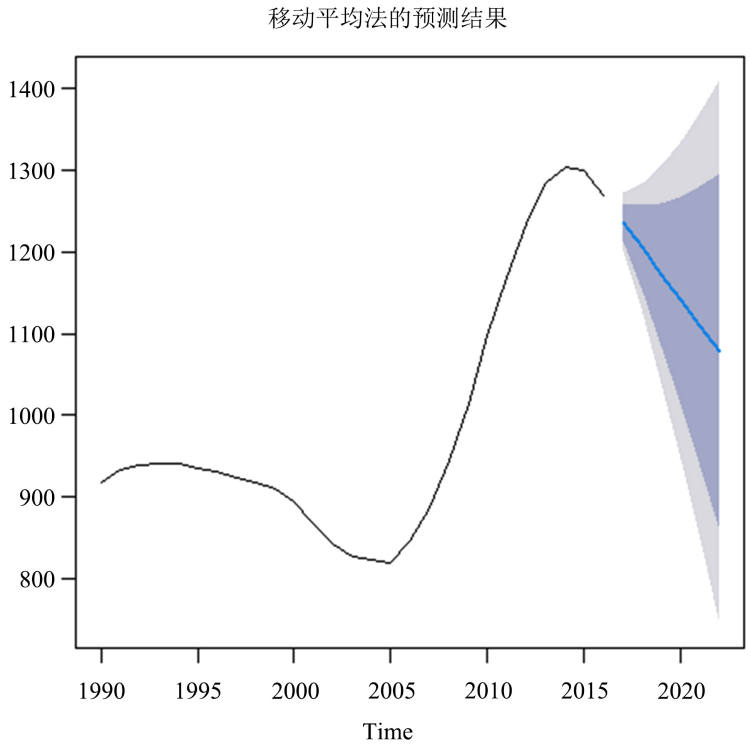

Figure 2. Forecast trend of moving average method
图 2. 移动平均法的预测趋势

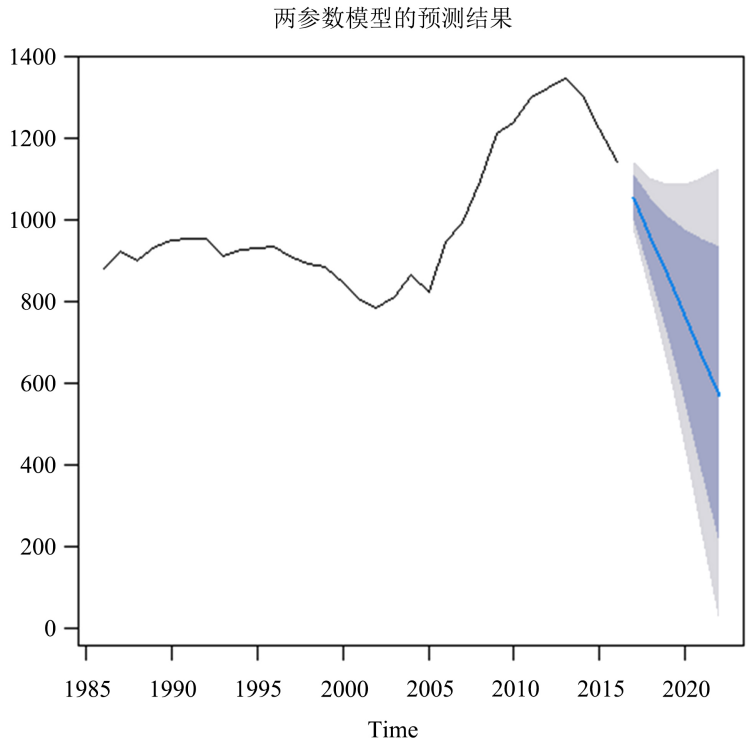

Figure 3. Predicted trends of the two parameters
图 3. 两参数的预测趋势

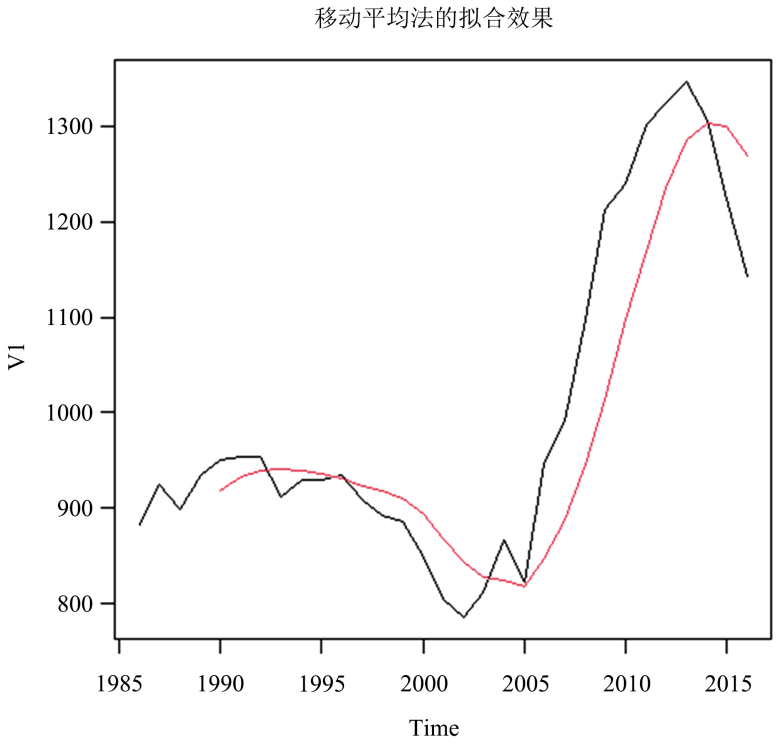

Figure 4. Fitting effect of moving average method
图 4. 移动平均法的拟合效果

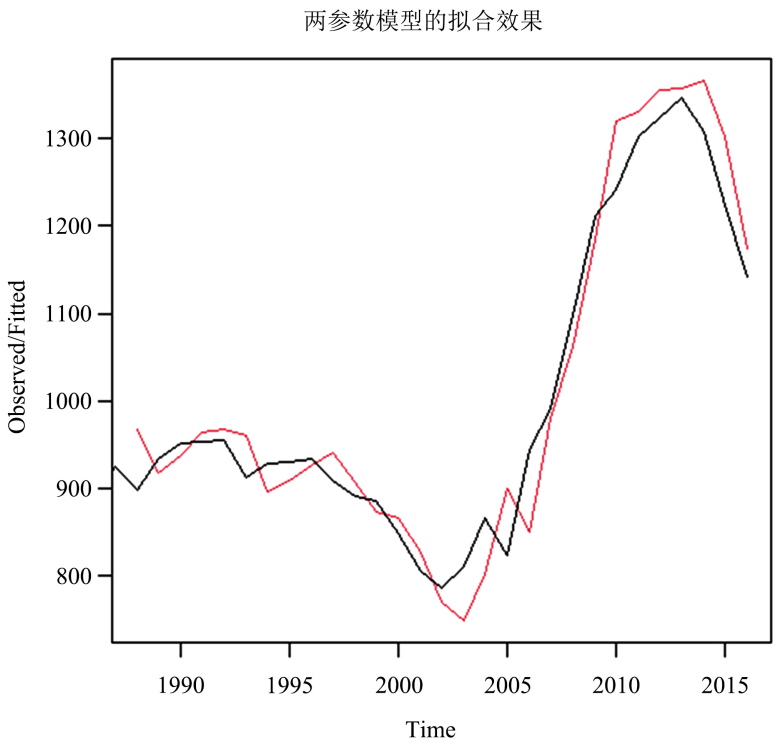

Figure 5. Fitting effect of the two-parameter model
图 5. 两参数模型的拟合效果

4. 结论

从以上结果可知，单纯利用灰色模型和时间序列模型只能预测数据变化的趋势，不能克服随机波动因素的影响。而改进的灰色－马尔可夫模型可以克服这一缺点，将预测精度提高。通过观察，在状态进行转化时，预测误差会增大，这可能与选取的状态有关，因此在使用灰色－马尔可夫模型时，要注意状态的选取。同时利用灰色－马尔可夫模型对 2022 年的数据进行了预测，得出 2022 年中国的结婚人口数据会有一定的增大。根据时间序列模型预测的 2022 年结婚人口数据没有可比性，误差很大。综上，选用灰色－马尔可夫模型可以较好地预测中国的结婚人口数据。

参考文献

[1] 王飞. 当代青年的婚恋观及其影响因素分析——基于 17-34 岁年龄段的青年调查数据[J]. 中国青年研究, 2015(7): 73-76, 81.

[2] 邓欣. 当前女大学生婚恋观状况及其影响因素研究[D]: [硕士学位论文]. 赣州: 江西理工大学, 2012.

[3] 陈兴广. 农村 70 后、80 后和 90 后婚恋观的特点及影响因素研究[D]: [硕士学位论文]. 北京: 中国青年政治学院, 2020.

[4] 黄桂仙. 新生代农民工婚恋观现状及其影响因素研究[D]: [硕士学位论文]. 昆明: 云南师范大学, 2015.

[5] 刘语佳. 中国石油消费预测模型研究与应用[D]: [硕士学位论文]. 北京: 北京交通大学, 2007.

[6] 翟维辉. 灰色马尔可夫组合预测模型的改进与应用[D]: [硕士学位论文]. 西安: 西安建筑科技大学, 2017.

[7] 吴倩. 中国石油消耗量的影响因素分析及其预测[D]: [硕士学位论文]. 湘潭: 湘潭大学, 2020.

[8] 王燕. 时间序列分析: 基于 R [M]. 北京: 中国人民大学出版社, 2015.

附录

```
y1<-read.csv("C:/Users/Administrator/Desktop/data1.csv",header=F)
y2<-read.csv("C:/Users/Administrator/Desktop/data2.csv",header=F)
y3<-read.csv("C:/Users/Administrator/Desktop/data3.csv",header=F)
y11<-ts(y1,start=c(1986,1),frequency =1)
y22<-ts(y2,start=c(1986,1),frequency =1)
y33<-ts(y3,start=c(1986,1),frequency =1)
plot(y11,main="黑色：原始值 绿色：GM(1,1) 蓝色：马尔科夫链")
lines(y22,col=3)
lines(y33,col=4)
yy<-read.csv("C:/Users/Administrator/Desktop/data.csv",header=F)###读取数据
X<-read.csv("C:/Users/Administrator/Desktop/测试集.csv",header=F)
yy1<-ts(yy,start=c(1986,1),frequency =1)
install.packages("tseries")
library(tseries)#---------------###单位根检验
adf.test(yy1)#-------------------####平稳的
plot(yy1,main="时序图")#-----------------------###时序图
acf(yy1,main="自相关图")
pacf(yy1,main="偏自相关图")
Box.test(yy1,type="Ljung-Box",lag=6)#---------###LB 统计量纯随机序列检验
Box.test(yy1,type="Box-Pierce",lag=6)#--------###Q 统计量纯随机序列检验
####平滑法（移动平均法）
install.packages("TTR")
library(TTR)
install.packages("forecast")
library(forecast)
y.fit2<-SMA(yy1,n=5)
plot(yy1,main="移动平均法的模拟效果")
lines(y.fit2,col=2)
y.fore2<-forecast(y.fit2,h=6)
y.fore2
plot(y.fore2,main="移动平均法的预测结果")
lines(X,col=3)###测试集
###两参数
y.fit3<-HoltWinters(yy1,gamma=F)
plot(y.fit3,main="两参数模型的模拟效果")
y.fore3<-forecast(y.fit3,h=6)
plot(y.fore3,main="两参数模型的模拟效果")
lines(yy1,col=3)
```