**Statistics and Application** 统计学与应用, 2021, 10(1), 151-161
Published Online February 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.101015

Hans汉斯

# 基于ARIMA模型的II型糖尿病患者数的建模与预测

**贾　雪，吴芷婧，孙佳萍，耿　帅，欧　圆，白晓东***

大连民族大学，辽宁 大连
Email: jiaxue688@126.com, *baixd518@126.com

收稿日期：2021年1月25日；录用日期：2021年2月19日；发布日期：2021年2月26日

## 摘　要

**目的：分析全国糖尿病疫情的时间分布特征，建立中国近年糖尿病时间序列分析的自回归移动平均模型(ARIMA)，预测病情未来发展趋势，为公众身体健康提出科学依据。方法：收集中国2000~2013年各年糖尿病患病人数数据，用R3.4.3软件构建ARIMA预测模型，对建立的模型进行参数估计、模型诊断，选择最优预测模型。利用构建的最佳模型对中国2014~2018各年糖尿病患病人数进行预测，并对预测效果进行评价。结果：ARIMA(1,1,0)模型为中国近年糖尿病人数的最优预测模型，其AIC、BIC的值分别为−38.93735、−37.80745，模型残差序列的Ljung-Box统计量 $\chi^2 = 12.408$，p值为0.4135，提示残差为白噪声序列，模型拟合良好。中国2014~2018糖尿病患病人数实际值与预测值的平均相对误差为2.27%，实际值均在预测值95%可信区间内。结论：ARIMA(1,1,0)模型能较好地模拟中国近年糖尿病患病人数的变化趋势，具有良好的预测效果。**

## 关键词

**糖尿病患病人数，最优模型，未来预测，发病趋势**

# Modeling and Prediction of the Number of Patients with Type II Diabetes Mellitus Based on ARIMA Model

**Xue Jia, Zhijing Wu, Jiaping Sun, Shuai Geng, Yuan Ou, Xiaodong Bai***

Dalian Minzu University, Dalian Liaoning
Email: jiaxue688@126.com, *baixd518@126.com

Received: Jan. 25th, 2021; accepted: Feb. 19th, 2021; published: Feb. 26th, 2021

*通讯作者。

文章引用: 贾雪, 吴芷婧, 孙佳萍, 耿帅, 欧圆, 白晓东. 基于 ARIMA 模型的 II 型糖尿病患者数的建模与预测[J]. 统计学与应用, 2021, 10(1): 151-161. DOI: 10.12677/sa.2021.101015

## Abstract

**Objective: To analyze the time distribution characteristics of diabetes in China, and to establish the autoregressive moving average model (ARIMA) for diabetes time series analysis in China in recent years, so as to predict the development trend of diabetes in the future and provide scientific basis for public health. Methods: The data of diabetes mellitus in China from 2000 to 2013 were collected and ARIMA prediction model was constructed by R3.4.3 software. The parameters of the model were estimated, the model was diagnosed, and the optimal prediction model was selected. The optimal model was used to predict the number of diabetic patients in China from 2014 to 2018, and the prediction effect was evaluated. Results: ARIMA (1,1,0) model was the best prediction model for the number of diabetes mellitus in China in recent years. The AIC and BIC values of ARIMA (1,1,0) were −38.93735 and −37.80745, respectively. Ljung box statistic of model residual sequence $\chi^2 = 12.408$, p value was 0.4135, indicating that the residual was white noise sequence, and the model fitted well. The average relative error between the actual value and the predicted value was 2.27%, and the actual value was within the 95% confidence interval of the predicted value. Conclusion: ARIMA (1, 1, 0) model can simulate the trend of diabetes mellitus in China in recent years, and has good prediction effect.**

## Keywords

**Diabetes Prevalence, Optimal Model, Future Prediction, Incidence Trend**



## 1. 引言

糖尿病是最常见的慢性病之一，随着人们生活水平的提高，人口老龄化以及肥胖发生率的增加，糖尿病的发病率呈逐年上升趋势。据统计，糖尿病在中国以每年 100 万的速度递增。其中，1 型糖尿病患者占 10%，2 型糖尿病患者占 90%。

根据 IDF 于 2017 年 11 月 14 日公布的数据显示：2012 年，全球半数以上糖尿病未被诊断，约 50%的糖尿病患者不知道自己患糖尿病。约 480 万人死于糖尿病，其中半数死于 60 岁以下。糖尿病防治医疗费用超过 4710 亿美元；2013 年，全球糖尿病在 20 岁~79 岁成人中的患病率为 8.3%，患者人数已达 3.82 亿，其中 80%在中等和低收入国家。IDF 预计到 2035 年，全球将有近 5.92 亿人患糖尿病。全球糖尿病防控形势已日趋严峻，糖尿病已对全球医疗体系构成巨大挑战。我国是糖尿病的重灾区，中国糖尿病的患病人数已高居全球首位，其次是印度、美国、巴西、俄罗斯。同时，据最新的《中国成人糖尿病流行与控制现状》调查研究显示，中国 18 岁及以上成人糖尿病患病率已高达 11.6%，糖尿病前期的患病率更是达到了惊人的 50.1%。这意味着，每 10 位中国成年人中，就有 6 位血糖不正常。按照这一比例，我国糖尿病患者人数已达 1.14 亿人，糖尿病前期人数接近 5 亿人。糖尿病已经成为我国最为重要和棘手的公共卫生问题之一[1]。本文利用 ARIMA 模型模拟中国近年糖尿病患者人数变化趋势，并进行短期预测，为中国人民身体健康提出科学依据。

## 2. 资料与方法

### 2.1. 数据来源

表 1 数据为搜集 2000~2018 年中国各年糖尿病人数。

### 2.2. 研究方法

ARIMA 模型是由 Box 和 Jenkings 提出的重要时间序列分析预测模型，ARIMA 模型包括三种基本模式：自回归模型 AR(p)，移动平均模型 MA(q)和自回归移动平均 ARMA(p,q)模型。其中，p、q 分别为自回归和移动平均的阶数。应用 ARIMA 模型的前提条件是：预测对象是一个零均值的平稳随机序列且这一平稳随机特性不随时间变化而变化[2]。如果序列不是平稳的序列，需要对原始序列进行平稳化的处理，最常见的处理方法是对原始序列进行 d 次差分，拟合预测模型 ARIMA(p,d,q)，d 代表非平稳序列差分的次数。本研究利用 R3.4.3 软件对中国近年糖尿病患者数据拟合 ARIMA 模型，并进行预测分析。ARIMA 模型建模分为 4 个步骤。

**1) 模型识别**

ARIMA 模型建模的基础前提是数据序列应该是平稳序列，本研究中通过观察中国近年来糖尿病患者序列的时序图，以此判断平稳性，若不满足平稳性，需要利用“diff”函数进行差分平稳化成平稳序列，此时差分的次数即为 ARIMA(p,d,q)模型中的阶数 d，在此基础上利用自相关函数图和偏自相关函数图判断截尾性和拖尾性，并且依据图像确定 p、q 的值进行定阶[3]。

**2) 参数估计**

采用最大似然估计计算自回归系数和移动平均系数。利用“Box.test”函数进行模型诊断，判断残差是否满足白噪声，若不满足白噪声模型则需要改进。

**3) 模型检验与优化**

依次确定 p、q 的值，利用“Box.test”函数进行模型诊断，判断残差是否满足白噪声，在残差满足白噪声前提下，依据信息量最小原则，选择 AIC 和 BIC 值最小的模型为最佳模型。

**4) 进行预测**

通过确定的最优模型，对中国 2014~2018 年糖尿病患者人数进行预测，并与实际值进行比较，计算预测值和实际值间的相对误差。

## 3. 结果

### 3.1. 模型识别

收集中国 2000~2018 年糖尿病患病人数的数据，见表 1，对中国 2000~2013 年糖尿病患病人数的数据绘制时序图，见图 1。由时序图观察可得，中国 2000 至 2013 年糖尿病患者人数整体呈上升趋势，2000 至 2005 年患者人数上升较快，2005 至 2010 年患者人数增长较为平缓，2010 至 2011 年最为特殊，患者人数出现了下降趋势，之后增长速度放缓。

由时序图可看出，此时间序列为非平稳时间序列，并且蕴含一个近似线性的递增趋势，因此，需要对该序列进行 1 阶差分运算，并绘制差分序列的时序图，见图 2。

由图 2 可见，经过一次差分之后，差分序列的线性趋势被提取。因此，对差分后的数据绘制 acf 图和 pacf 图，见图 3。

由图 3 我们可以看出，自相关函数图呈拖尾现象，偏自相关函数图呈现了 1 阶截尾特性，因此，我

**Table 1.** Number of diabetes mellitus patients in China from 2000 to 2018 (unit: billion)
**表 1.** 2000~2018 年中国糖尿病患病人数(单位：亿)

| 时间(年) | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 |
|---|---|---|---|---|---|---|---|---|---|
| | 0.1055 | 0.1326 | 0.1742 | 0.238 | 0.3827 | 0.5584 | 0.6401 | 0.7135 | 0.7918 |
| 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 |
| 0.8847 | 0.92 | 0.9 | 0.924 | 0.984 | 0.9629 | 1.0965 | 1.1 | 1.144 | 1.149 |

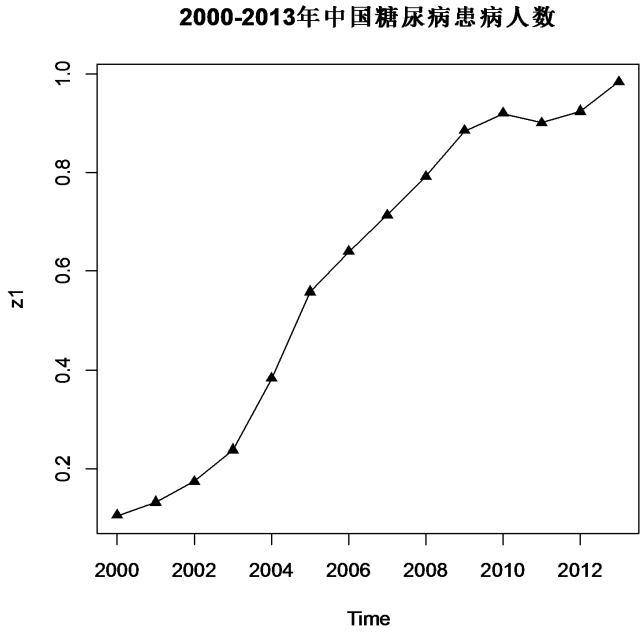


**Figure 1.** Number of diabetes patients in China from 2000 to 2013
**图 1.** 2000~2013 年中国糖尿病患病人数

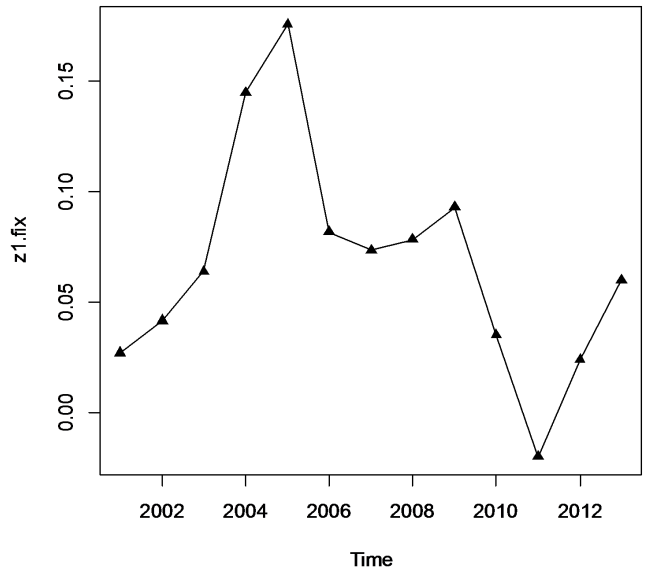


**Figure 2.** Sequence Diagram of first order difference sequence
**图 2.** 一阶差分序列时序图

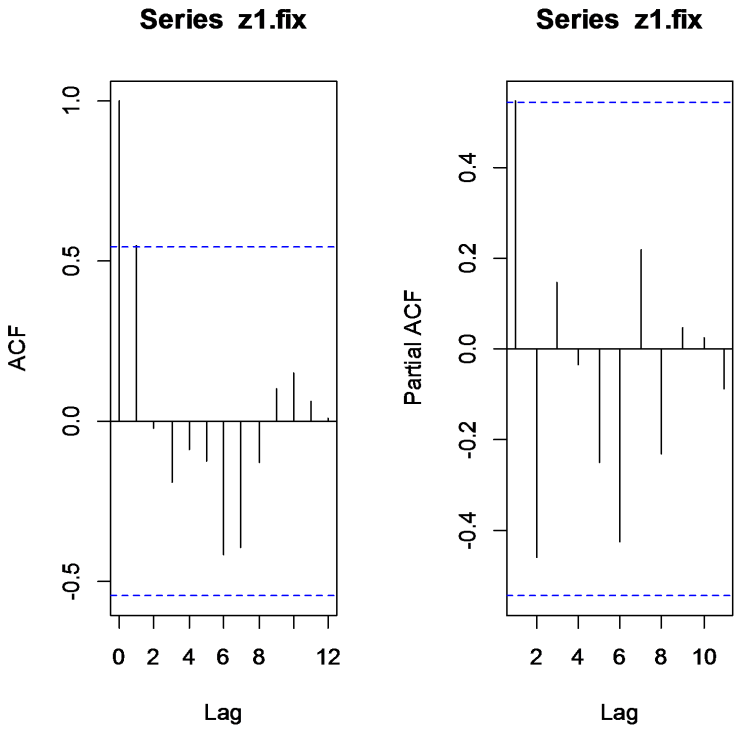


**Figure 3.** ACF Diagram and pacf diagram of difference sequence
**图 3.** 差分序列的 acf 图和 pacf 图

们可以初步确定拟合模型为 ARIMA(1,1,0)。

### 3.2. 参数估计

由 3.1 我们已经初步将模型识别为 ARIMA(1,1,0)，现采用条件最小二乘法估计未知参数，根据估计的结果，我们可以确定该模型的口径为

$$x_t = 0.8622x_{t-1} + \varepsilon_t$$

接下来，我们利用“Box.test”函数进行检验，对中国近年糖尿病患者序列拟合模型的残差序列进行延迟 6 阶和延迟 12 阶的白噪声检验，检验结果见表 2。

**Table 2.** Residual test results of Arima (1,1,0) model
**表 2.** ARIMA(1,1,0)模型的残差检验结果

| 延迟阶数 | $\chi^2$ 值 | P 值 |
|---|---|---|
| 6 | 9.9583 | 0.1264 |
| 12 | 13.852 | 0.3102 |

由表 2，我们可以知道，延迟 6 阶和延迟 12 阶的统计量 $Q_{LB}$ 的 p 值显著大于显著性水平 0.05，所以可以认为模型的残差序列是白噪声序列。

### 3.3. 模型检验与优化

我们依次将模型拟定为 ARIMA(1,1,0)、ARIMA(0,1,1)、ARIMA(0,1,2)、ARIMA(2,1,0)、ARIMA(1,1,1)、

ARIMA(2,1,1)、ARIMA(1,1,2)，然后利用“Box.test”函数进行模型诊断，判断残差是否满足白噪声，检验结果见表 3。

**Table 3.** White Noise test for residual series of Arima (P, 1, Q) alternative model
**表 3.** ARIMA(p,1,q)备选模型残差序列的白噪声检验

| | df = 6 | | df = 12 | |
|---|---|---|---|---|
| | $\chi^2$ | p 值 | $\chi^2$ | p 值 |
| ARIMA(1,1,0) | 8.9688 | 0.1753 | 12.408 | 0.4135 |
| ARIMA(0,1,1) | 3.5783 | 0.7335 | 7.7999 | 0.8006 |
| ARIMA(0,1,2) | 4.8535 | 0.5627 | 6.8126 | 0.8697 |
| ARIMA(2,1,0) | 8.7234 | 0.1897 | 11.482 | 0.4881 |
| ARIMA(1,1,1) | 5.6635 | 0.4619 | 7.3736 | 0.832 |
| ARIMA(2,1,1) | 5.2502 | 0.5121 | 7.1655 | 0.8465 |
| ARIMA(1,1,2) | 3.9477 | 0.6838 | 6.0152 | 0.9153 |

由表 3 可以知道，7 个模型延迟 6 阶和延迟 12 阶的统计量 $Q_{LB}$ 的 p 值均显著大于显著性水平 0.05，因此可以认为这 7 个模型的残差序列均是白噪声序列。由于残差序列均满足白噪声，因此依据信息量最小原则，我们需要选择 AIC 和 BIC 值最小的模型为最佳模型。7 个模型的 AIC、BIC 值见表 4。

**Table 4.** Comparison of Arima (P, 1, Q) alternative model AIC and BIC values
**表 4.** ARIMA(p,1,q)备选模型 AIC、BIC 值的比较

| 模型 | AIC | BIC |
|---|---|---|
| ARIMA(1,1,0) | −38.93735 | −37.80745 |
| ARIMA(0,1,1) | −35.05312 | −33.92322 |
| ARIMA(0,1,2) | −36.57639 | −34.88154 |
| ARIMA(2,1,0) | −37.56978 | −35.87493 |
| ARIMA(1,1,1) | −38.76389 | −37.06904 |
| ARIMA(2,1,1) | −36.98161 | −34.72181 |
| ARIMA(1,1,2) | −37.42787 | −35.16807 |

依据表 4，可以看出 ARIMA(1,1,0)的 AIC、BIC 值最小。因此，我们选定 ARIMA(1,1,0)为最优模型。

### 3.4. 进行预测

通过 3.3 中，我们确定最优模型为 ARIMA(1,1,0)。因此，我们通过最优模型进行未来五年，即 2014~2018 年中国糖尿病患者人数的预测，并绘制图像，图像见图 4。

并绘制个性化的预测图，预测图见图 5。

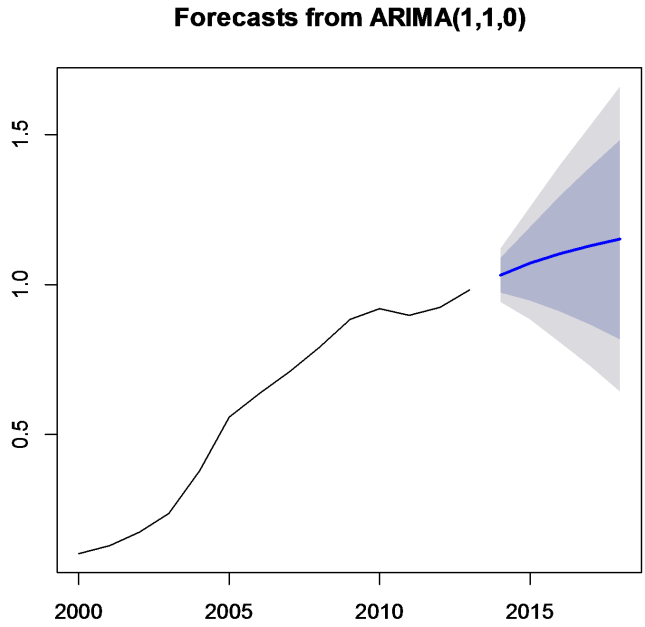


**Figure 4.** Projected number of diabetes patients in China from 2014 to 2018
**图 4.** 2014~2018 年中国糖尿病患者人数预测图

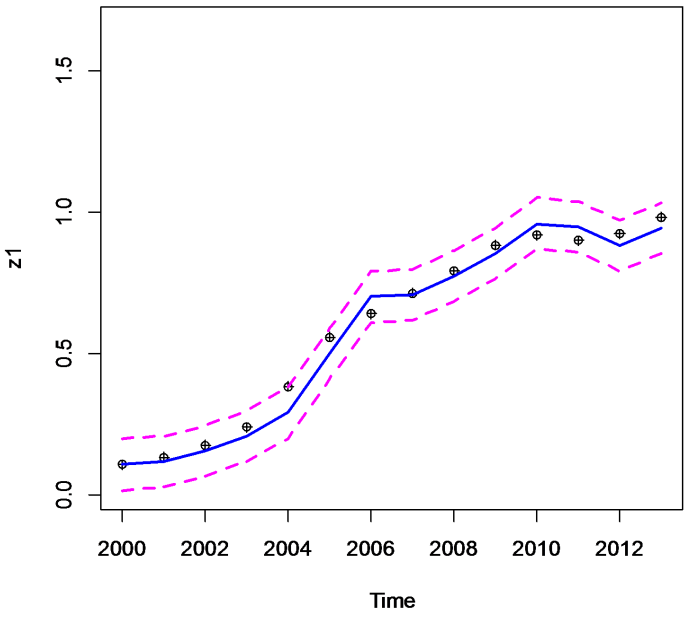


**Figure 5.** Personalized predictors of Diabetes Mellitus in China from 2014 to 2018
**图 5.** 2014~2018 年中国糖尿病患者人数个性化预测图

散点图为观察值序列，实线是拟合值，虚线是 95%的置信线。

同时将预测值与这五年间糖尿病患者人数的实际值进行比较，计算相对误差，判断预测效果。结果见表 5。

由表 5 我们发现，实际糖尿病患病人数均在预测值 95%可信区间内，用 2014~2018 中国糖尿病实际患病人数评价 ARIMA(1,1,0)模型预测准确度，预测值与实际值的平均相对误差为 2.2689%，可见预测结果与实际结果基本一致，模型预测效果较好[4]。

**Table 5.** Forecast analysis of Arima (1,1,0) model
**表 5.** ARIMA(1,1,0)模型预测分析

| 日期 | 实际值 | 预测值 | 95%可信区间 | 误差 | 相对误差(%) |
|---|---|---|---|---|---|
| 2014 | 0.9629 | 1.0329 | 0.9421~1.1236 | 0.07 | 7.2697 |
| 2015 | 1.0965 | 1.0726 | 0.8846~1.2607 | 0.0239 | 2.1797 |
| 2016 | 1.1 | 1.1050 | 0.8119~1.3982 | 0.005 | 0.4545 |
| 2017 | 1.144 | 1.1314 | 0.7302~1.5326 | 0.0126 | 1.1014 |
| 2018 | 1.149 | 1.1529 | 0.6435~1.6622 | 0.0039 | 0.3394 |

## 4. 总结

近年来，由于生活水平的提高，饮食结构的改变，日趋紧张的生活节奏以及少动多坐的生活方式等诸多因素，全球糖尿病发病率增长迅速，糖尿病已经成为继肿瘤、心血管病变之后第三大严重威胁人类健康的慢性疾病，未来 50 年内，糖尿病仍将是中国一个严重的公共卫生问题[5]。专家指出，由于目前人们的饮食结构正在由植物型向动物型转变，高脂肪、高热量食物正在越来越多地充斥我们的生活，加上糖尿病知识以及健康生活理念不够普及，都对中国糖尿病防治能力以及糖尿病教育提出更高的要求。尽管中国糖尿病发病趋势严峻，但防治状况不容乐观，一是因为目前我国专业糖尿病治疗机构、人员和设备等资源不足，无法与日益增长的糖尿病患者人数相适应，致使中国整体糖尿病诊治率还相对较低，二是因为许多公众和患者对糖尿病防治知识的认识不足，在糖尿病防治方面存在治疗不及时、用药选择和时机不当、忽视饮食运动等误区，从而使治疗效果不理想。另外，我国在糖尿病营养学和运动领域几乎还是空白，绝大多数医院目前尚无专业的糖尿病营养和运动医师[6]。在糖尿病防治过程中，护理人员的专业素养与西方发达国家相比相对滞后，护理人员糖尿病专业知识的掌握程度还比较肤浅。同时，由于护理人力资源配置不合格，无暇对糖尿病人进行教育。

时间序列分析通过观察值历史数据来预测其发展趋势，探索其时间分布特征及影响因素。ARIMA 模型应用于经济、工程、气象、水利等众多领域。医学卫生领域也用于预测疾病发病率、人群寿命、医疗卫生费用和食物中毒等，还用于研究气候及环境因素对人群健康的影响[7]。

本研究选用中国 2000~2013 年糖尿病患者人数的数据进行时间序列分析。显示这些年间中国糖尿病患者人数整体呈上升趋势，且前些年大幅上升。基于我国基本国情分析，自 2000 年起，中国经济发展迅速，人民生活水平日益提高，人民的饮食结构正在由植物型向动物型转变，有更多高脂肪的食物经常出现在人们的生活中，因此，这些年间，糖尿病患病人数大幅增长。2010 年后，糖尿病患者人数增长放缓，由于前些年人们对于糖尿病有了一定的认识，且我国的医疗水平日益提高，因此，对于糖尿病的防治有了一定的效果，增长速度放缓。经过模型识别、参数估计、模型优化确定最优模型为 ARIMA(1,1,0)，并且通过最优模型我们对中国 2014~2018 这五年间糖尿病患者人数进行了预测，预测值与实际值的平均相对误差为 2.2689%，且实际糖尿病患病人数均在预测值 95%置信区间内，预测效果较好。通过预测，我们发现这五年间，中国糖尿病患者人数仍然是逐年上升，但增速较为缓慢，可见，随着信息化时代的发展，人们对于糖尿病的认识越来越广泛，对于糖尿病的防治越来越有效，人们通过健康的生活方式，并且在日常生活中，控制饮食，加强锻炼，以此来预防糖尿病的发生。

本研究证实，通过时间序列对某种疾病患病人数或发病率进行建模，确定最优模型，可较好地对这种疾病的发展趋势进行预测，这种方法给疾病的预测及控制提供了一种科学的依据。

## 基金项目

国家级大创项目(202012026039)。

## 参考文献

[1] 郭启煜. 扑面而来的高糖时代没人能置身事外[J]. 养生大世界, 2018(1): 20-23, 25.

[2] 马翠荣, 杨婕, 余小金. 江苏省 2006-2014 年城乡未成年人跌倒病例的时间序列预测分析[J]. 中华疾病控制杂志, 2018, 22(2): 122-125, 137.

[3] 严宙宁, 牟敬锋, 赵星, 严燕, 罗文亮. 基于 ARIMA 模型的深圳市大气 PM2.5 浓度时间序列预测分析[J]. 现代预防医学, 2018(2): 220-223, 242.

[4] 王媛媛, 田飞, 刘晶磊. 时间序列分析在北京市东城区艾滋病病毒感染者和艾滋病患者发病率预测中的应用[J]. 疾病监测, 2017(9): 731-734.

[5] 周行, 夏木, 杜宇. 中国顶级糖尿病专家告诉您 怎么应对“甜蜜的负担”[J]. 养生大世界, 2018(1):20-23.

[6] 糖尿病毕业论文终稿[EB/OL]. https://www.docin.com/p-2228527251.html&dpage=1&key=%E7%B3%96%E5%B0%BF%E7%97%85%E6%80%8E%E4%B9%88%E6%B2%BB&isPay=-1&toflash=0&toImg=0

[7] 任江萍, 陈直平, 孙继民, 陈恩富, 施旭光, 张蓉, 刘营, 凌锋. 全国人间狂犬病疫情的时间序列分析[J]. 中国人兽共患病学报, 2018, 34(3): 239-242.

## 附录

**中国近年糖尿病患者数据建模的程序**

**1) 模型识别**

```
> x<-read.csv("D:/时间序列/中国历年糖尿病患病人数.csv",header=T)
> z1<-ts(x$data,start=2000)
> plot(z1,type="o",pch=17,main="2000-2013 年中国糖尿病患病人数")  #绘制时序图
> z1.fix<-diff(z1)
> plot(z1.fix,type="o",pch=17)  #消除趋势
> par(mfrow=c(1,2))
> acf(z1.fix,lag=50)
> pacf(z1.fix)    #绘制自相关、偏自相关图
```

**2) 参数估计、模型检验**

```
> l<-arima(z1,order=c(1,1,0),method="CSS")
> l
> for(i in 1:2) print(Box.test(l$residual,type="Ljung-Box",lag=6*i))
```

**3) 模型优化**

```
> library(forecast)
> z1.fix1<-Arima(z1,order=c(1,1,0),method="ML")
>for(i in 1:2) print(Box.test(z1.fix1$residual,type="Ljung-Box",lag=6*i))
> z2.fix2<-Arima(z1,order=c(0,1,1),method="ML")
>for(i in 1:2) print(Box.test(z2.fix2$residual,type="Ljung-Box",lag=6*i))
> z3.fix3<-Arima(z1,order=c(0,1,2),method="ML")
>for(i in 1:2) print(Box.test(z3.fix3$residual,type="Ljung-Box",lag=6*i))
> z4.fix4<-Arima(z1,order=c(2,1,0),method="ML")
>for(i in 1:2) print(Box.test(z4.fix4$residual,type="Ljung-Box",lag=6*i))
> z5.fix5<-Arima(z1,order=c(1,1,1),method="ML")
>for(i in 1:2) print(Box.test(z5.fix5$residual,type="Ljung-Box",lag=6*i))
> z6.fix6<-Arima(z1,order=c(2,1,1),method="ML")
>for(i in 1:2) print(Box.test(z6.fix6$residual,type="Ljung-Box",lag=6*i))
> z7.fix7<-Arima(z1,order=c(1,1,2),method="ML")
>for(i in 1:2) print(Box.test(z7.fix7$residual,type="Ljung-Box",lag=6*i))
> z1.fix1$aic
> z1.fix1$bic
> z2.fix2$aic
> z2.fix2$bic
> z3.fix3$aic
> z3.fix3$bic
> z4.fix4$aic
> z4.fix4$bic
```

```
> z5.fix5$aic
> z5.fix5$bic
> z6.fix6$aic
> z6.fix6$bic
> z7.fix7$aic
> z7.fix7$bic
```

**4) 进行预测**

```
> z1.fore<-forecast(z1.fix1,h=5)
> z1.fore
> plot(z1.fore)     #模型预测，进行五期预测
```

绘制个性化预测图：

```
> Q1<-z1.fore$fitted-1.96*sqrt(z1.fix1$sigma2)
> Q2<-z1.fore$fitted+1.96*sqrt(z1.fix1$sigma2)
> B1<-ts(z1.fore$lower[,2])
> B2<-ts(z1.fore$upper[,2])
> J1<-min(z1,Q1,B1)
> J2<-max(z1,Q2,B2)
> plot(z1,type="p",pch=10,ylim=c(J1,J2))
> lines(z1.fore$fitted,col=4,lwd=2)
> lines(z1.fore$mean,col=4,lwd=2)
> lines(Q1,col=6,lty=2,lwd=2)
> lines(Q2,col=6,lty=2,lwd=2)
> lines(B1,col=7,lty=2,lwd=2)
> lines(B2,col=7,lty=2,lwd=2)   #绘制个性化预测图
```