Advances in Social Sciences 社会科学前沿, 2019, 8(1), 69-74
Published Online January 2019 in Hans. http://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2019.81011

Hans汉斯

Population Quantity Model of Yunnan Province Based on Time Series Analysis

Jianying Yang

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan
Email: 346290100@qq.com

Received: Dec. 28th, 2018; accepted: Jan. 10th, 2019; published: Jan. 17th, 2019

Abstract

Population problem is a huge problem, which is related to many aspects of society. In view of the fact that the population of Yunnan Province is greatly affected by the previous period and its growth is in a linear way, this paper establishes ARIMA model and linear model, and takes mean square error (*RMSE*) and average absolute percentage error (*MAPE*) as the accuracy evaluation indexes, and finally draws the conclusion that the ARIMA model has the highest accuracy. Based on this model, we predict the population of Yunnan Province in the next five years, and get the conclusion that the population of Yunnan Province will increase steadily.

Keywords

Population Mode, Linear Mode, ARIMA Mode, Accuracy Evaluation

基于时间序列分析的云南省人口数量模型

杨健颖

云南财经大学，统计与数学学院，云南 昆明
Email: 346290100@qq.com

收稿日期：2018年12月28日；录用日期：2019年1月10日；发布日期：2019年1月17日

摘 要

人口问题是一个巨大的问题，这关系到社会的方方面面。本文针对云南省人口当期受前期影响大，且增长呈现出接近线性方式的特点，建立了ARIMA模型和线性模型，以均方误差(*RMSE*)和平均绝对百分比误差(*MAPE*)作为精度评价指标，最终得出ARIMA模型精度最高的结论。并以此模型对云南省未来5年的人口数量进行预测，得出云南省人口将会稳定增长的结论。

文章引用: 杨健颖. 基于时间序列分析的云南省人口数量模型[J]. 社会科学前沿, 2019, 8(1): 69-74.
DOI: 10.12677/ass.2019.81011

关键词

人口模型，线性模型，ARIMA模型，精度评价

1. 引言

人口问题是一个巨大的社会性问题，一个地区人口的发展会影响到该地区的社会、经济和环境资源等各个方面，因此，对人口的有效预测能为将来制定合理的发展方案提供有力依据。

云南省是人口大省，作为国家东南亚和南亚开放的门户，在一带一路的国家战略带来的机遇下，对云南省人口的有效的预测能为其制定合理的发展方案提供有力依据，因此，建立一个精度高的云南省人口数量模型就显得尤为重要了。

预测人口应该结合人口的发展特点来进行，由于人口数量当前受前期影响较为大，而通过简单分析能得出云南省近 45 年来人口呈现接近线性方式增长的结论。结合这些特点，本文旨在通过构建一元线性模型和 ARIMA 模型来对云南省人口数量模型进行拟合，并从中选取精度最高的模型作为云南省人口数量模型，以此来预测云南省的人口数量的发展。

2. 文献综述

人口问题是一个世界性问题，这影响到社会的方方面面，从而对人口的有效预测也就成了一个非常重要的问题。田飞[1]对人口预测的方法体系进行的研究，总结出人口预测模型发展的历程：较早期的模型有指数增长模型、Logistic 模型和线性回归模型，随着模糊数学和统计学的发展出现了灰色系统 GM (1,1) 模型和时间序列模型，而生物医学和计算机的发展为此带来了神经网络算法[2]，这些算法各有优缺点，在模型的选取上应该结合不同数据的特点选择合适的模型。

人口的预测模型应结合当地人口发展的特点来选取。云南省的人口特点除了包含当期人口规模受前期人口影响较大这个几乎所有人口都有的特点以外，还呈现接近线性方式增长的特点，第一个适用于 ARIMA 模型，第二个适用于线性趋势模型。这两个模型应该都十分广泛，其中第二个模型就是普通的以时间序号为自变量的一元线性模型，易于理解不做过多介绍。

时间序列模型在分析前期对当期有影响的数据时具有非常广泛的应用。如李子奈等[3]使用时间序列模型分析和预测了我国的通货膨胀情况，结果表明时间序列模型的精度高。张松林等[4]使用时间序列模型分析我国的城市化水平，并给出高精确度的预测结果；唐毅[5]在传统时间序列模型的基础上，从样本序列的动态选取及模型识别两方面进行优化，提出了一种能动态调整模型参数的改进时间序列模型。而在近两年的研究中，时间序列的应用有了更广泛的实际应用，如王莉等[6]使用时间序列模型结合残差控制图分析了兰州市的空气质量指数，证实了将时间序列模型与残差控制图结合预测监控大气污染的有效性。刘自强等[7]运用关键词群分析、社会网络分析和时间序列模型分析预测其研究热点的发展趋势，结果表明该方法的有效性。这也进一步表明了传统的时间序列模型与新方法的结合能有效克服传统时间序列模型缺陷，有效解决时代发展的新问题。

云南省人口发展当前受前期影响大，且呈现接近线性方式增长。正是基于以上分析，我们分别使用ARIMA模型和线性模型来对云南省人口数量模型进行拟合，并从中选取精度最高的模型作为云南省人口数量模型，以此来预测云南省的人口数量的发展。

3. 理论知识

3.1. 趋势拟合法——线性模型[8]

趋势拟合法用于拟合具有趋势项的数据，把时间作为自变量，相应的序列观察值作为因变量，建立序列值随时间变化的回归模型的方法。而当数据呈现线性或接近线性方式增长时，使用线性模型来拟合即可。此时模型表达式为：

$$\begin{cases} x_t = a + bt + I_t \\ E(I_t) = 0, Var(I_t) = \delta^2 \end{cases}$$

其中 $\{I_t\}$ 为随机波动，$T_t = a + bt$ 为消除随机波动的影响之后该序列的长期趋势。

3.2. ARIMA 模型[9]

具有如下结构的模型称为求和自回归移动平均模型，简记为 $ARIMA(p,d,q)$ 模型：

$$\begin{cases} \Phi(B)\nabla^d x_t = \Theta(B)\varepsilon_t \\ E(\varepsilon_t) = 0, Var(\varepsilon_t) = \delta_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ E(x_s \varepsilon_t) = 0, \forall s \neq t \end{cases}$$

式中，$\nabla^d = (1-B)^d$；$\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$，为平稳可逆 $ARMA(p,q)$ 模型的自回归系数多项式；$\Theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$，为平稳可逆 $ARMA(p,q)$ 模型的移动平滑系数多项式。

3.3. 精确度评价

本文使用 $(x_1, x_2, \cdots x_n)$ 表示实际值，$(x_1, x_2, \cdots x_n)$ 表示实际值表示预测值，n 为样本数。

1) 均方误差($RMSE$)：

$$RMSE = \left(\frac{1}{n} \sum_{i=1}^{n} \left(x_i - \widehat{x_i} \right)^2 \right)^{1/2}$$

2) 平均绝对百分比误差($MAPE$)：

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{x_i - \widehat{x_i}}{x_i} \right| \times 100\%$$

显然，均方误差($RMSE$)和平均绝对百分比误差($MAPE$)越小，说明模型预测精确度越高，误差越小，模型拟合效果也就越好。

4. 模型构建

本文使用 R 语言[10]进行模型的构建。本文选取 1973 年至 2017 年共 45 年的云南人口数量数据(本数据来源于《云南省统计年鉴》，单位为万，用 x 表示)，画出其趋势图见图 1。

从图 1 可得出，云南省人口数量一直呈现递增的趋势，且增长趋势接近线性趋势，从而使用趋势拟合法中的线性模型来进行拟合。

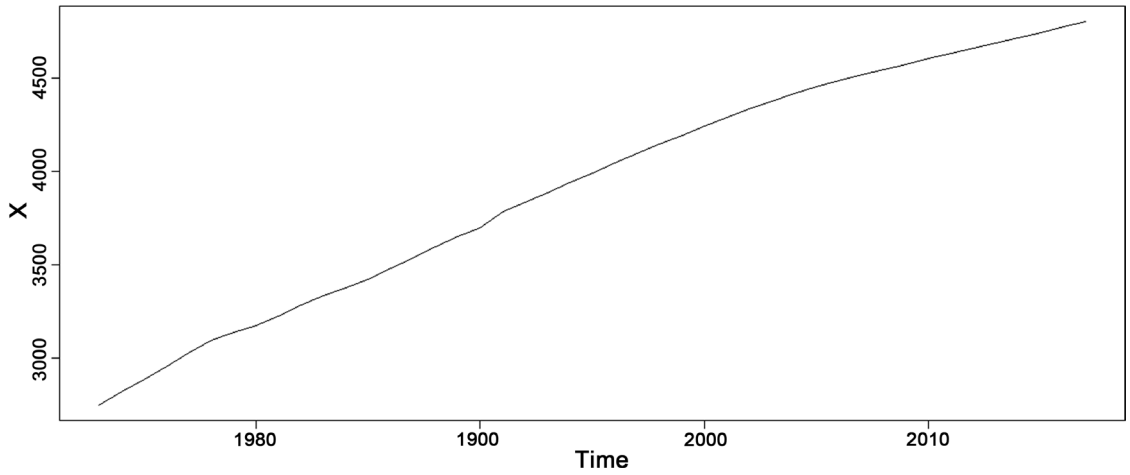

Figure 1. Population trend map of Yunnan Province in the past years
图 1. 云南省历年人口趋势图

4.1. 趋势拟合法——线性模型

1) 模型拟合

此模型是时间作为自变量，相应观察值作为因变量的一元线性模型，得出模型拟合结果(t 表示作为自变量的时间)：

$$x = 2820.9909 + 47.5129t \tag{1}$$

$$F = 3396,\ R^2 = 0.9872$$

2) 模型检验

模型(1)的 $p = 2.2\times10^{-16}$；常数项的 $p = 2\times10^{-16}$；系数 t 的 $p = 2\times10^{-16}$，此三者均小于显著性水平 $\alpha = 0.05$，表示模型(1)及其系数 t 均显著，模型(1)合理。

3) 精度评价

依据模型(1)可以计算通过此模型得出的云南省1973年至2015年人口数量的估计值，$RMSE$ 和 $MAPE$，得：

$$RMSE = 69.44,\ MAPE = 1.46\%$$

4.2. ARIMA 模型[8]

1) 数据预处理

通过图 1 中显示的人口数量递增趋势可知 x 为非平稳序列，我们对其进行逐步差分，并对逐步差分的结果进行 DF 检验，以判断逐步差分后序列的平稳性。

一阶差分 $P = 0.5289$；二阶差分 $P = 0.01$，则二阶差分后序列(用 x_1 表示)能通过平稳性检验，本文选用差分两次的序列来拟合 ARIMA 模型。

2) 模型拟合

我们这里画出 x_1 的自相关和偏自相关图，以判断 ARIMA 模型中的自回归和移动平均的阶数。该图见图 2。

从图 2 可以看出，1 期的 ACF 值在二倍标准差之外；5 期的 ACF 值在刚好等于二倍标准差；1 期的 PACF 值在二倍标准差之外，从而我们尝试拟合包含一阶自回归、一阶移动平均和一阶自回归、五阶移动平均的 ARIMA 模型。

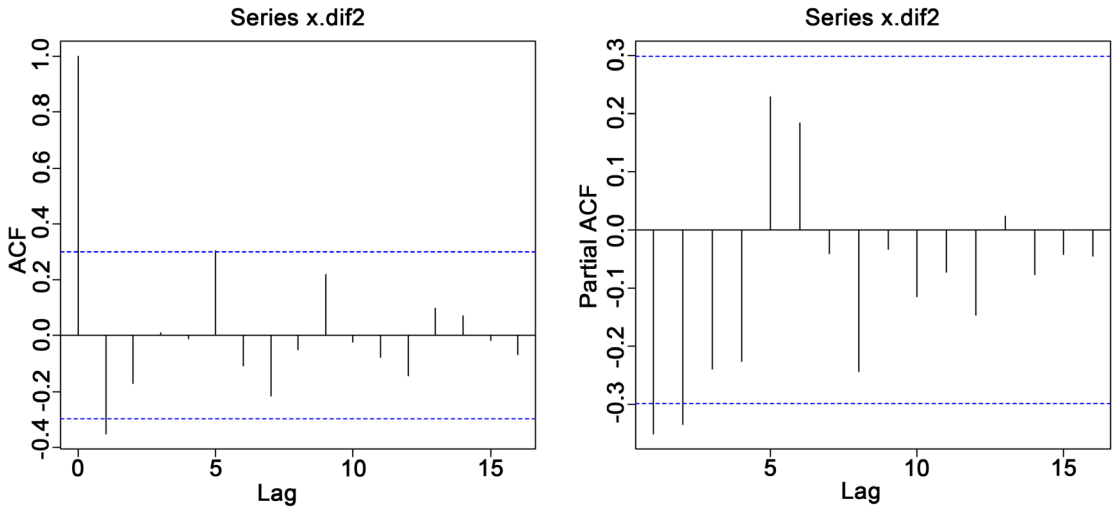

Figure 2. Autocorrelation diagram and partial autocorrelation diagram of x_1
图 2. x_1 的自相关和偏自相关图

我们拟合的这两种模型的 AIC 值分别为 315.41 和 315.82，以 AIC 值最小准则选取一阶自回归、一阶移动平均的 ARIMA 模型较为合理。从而拟合的模型为：

$$x_t = 0.0731x_{t-1} + \varepsilon_t - 0.5731\varepsilon_{t-1} \tag{2}$$

3) 模型检验

我们这里对模型(2)进行纯随机性检验，6 期和 12 期检验结果的 P 值分别为 0.3365 和 0.1915，此二者均大于显著性水平 $\alpha = 0.05$，表明模型(2)通过残差的随机性检验，模型(2)是合理的。

4)精度评价

依据模型(2)可以计算通过此模型得出的云南省 1973 年至 2015 年人口数量的估计值，并以此计算 *RMSE* 和 *MAPE*，得：

$$RMSE = 8.62,\ MAPE = 0.15\%$$

4.3. 模型比较

我们以均方误差(*RMSE*)和平均绝对百分比误差(*MAPE*)作为评价，模型精确度的指标，此二者值越小说明模型拟合效果越好。模型(1)和模型(2)的精度比较见表 1。

Table 1. Precision comparisons of Model (1) and Model (2)
表 1. 模型(1)和模型(2)的精度比较表

RMSE		*MAPE*	
模型(1)	模型(2)	模型(1)	模型(2)
69.44	8.62	1.46%	0.15%

从表 1 可以看出，ARIMA 模型的 *RMSE* 值和 *MAPE* 值两个指标都比线性模型的小，这表明 ARIMA 模型精度都大于线性模型，从而我们选取模型(2)作为云南省人口数量的预测模型。

5. 模型预测

我们得出的云南省人口预测模型为：

$$x_t = 0.0731x_{t-1} + \varepsilon_t - 0.5731\varepsilon_{t-1} \quad (2)$$

使用 R 语言对模型(2)对云南省今后 5 年的人口数据进行预测，得出的预测值及其 95%置信区间间表 2。

Table 2. Population forecast value table of Yunnan Province in the next five years
表 2. 云南省今后 5 年人口数量预测值表

年份	预测值(万)	95%置信区间
2018	4828.7	(4811.447, 4845.958)
2019	4857.38	(4826.277, 4888.49)
2020	4886.06	(4840.07, 4932.055)
2021	4914.74	(4852.573, 4976.91)
2022	4943.42	(4863.782, 5023.059)

我们这里计算真实的云南省 2013 年~2017 年人口增长率为 22.6，预测的 2018 年~2022 年的人口增长率为 22.944，此二者值非常接近。这说明云南省未来五年的人口将会基本保持目前速度。

6. 结论

本文根据云南省人口当期受前期影响大，且增长呈现接近线性趋势的特点分别构建了 ARIMA 模型和线性模型，以均方误差(*RMSE*)和平均绝对百分比误差(*MAPE*)作为评价精度的指标，结果表明 ARIMA 模型的精度最高，最终选取 ARIMA 模型作为云南省人口数量模型。

本文以 ARIMA 模型预测云南省未来五年的人口数量，预测结果表明云南省未来五年人口将保持稳定增长，且预测未来五年(2018~2022 年段)人口数量曲线的斜率与 2013~2017 年段曲线斜率非常接近，这说明增长速度与目前增长速度基本相同。

基金项目

本文得到了云南财经大学研究生创新基金项目(2018YUFEYC001)的资助。

参考文献

[1] 田飞. 人口预测方法体系研究[J]. 安徽大学学报(哲学社会科学版), 2011, 35(5): 151-156.

[2] 尹春华, 陈雷. 基于 BP 神经网络人口预测模型的研究与应用[J]. 人口学刊, 2005(2): 44-48.

[3] 叶阿忠, 李子奈. 我国通货膨胀的混合回归和时间序列模型[J]. 系统工程理论与实践, 2000, 20(9): 138-140.

[4] 张松林, 熊红轶. 中国城市化水平时间序列模型分析: 1949~2007 [J]. 统计与决策, 2009(20): 80-82.

[5] 唐毅, 刘卫宁, 孙棣华, 等. 改进时间序列模型在高速公路短时交通流量预测中的应用[J]. 计算机应用研究, 2015, 32(1): 146-149.

[6] 王莉, 赵渊, 杨显明, 等. 基于时间序列模型与残差控制图的兰州市空气质量研究[J]. 高原气象, 2015, 34(1): 230-236.

[7] 刘自强, 王效岳, 白如江. 基于时间序列模型的研究热点分析预测方法研究[J]. 情报理论与实践, 2016, 39(5): 27-33.

[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.

[9] 潘省初. 计量经济学中级教程[M]. 第 2 版. 北京: 清华大学出版社, 2013.

[10] 吴喜之. 应用时间序列分析-R 软件陪同[M]. 北京: 机械工业出版社, 2014.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2169-2556，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：ass@hanspub.org