Statistics and Application 统计学与应用, 2024, 13(6), 2375-2381
Published Online December 2024 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2024.136230

Hans汉斯

基于ARIMA模型的河北省GDP预测研究

马　涛[1]，于　涧[1]，于泽翔[2]，刘红星[1]

[1]沈阳师范大学数学与系统科学学院，辽宁　沈阳
[2]东北大学悉尼智能科技学院，辽宁　沈阳

收稿日期：2024年11月20日；录用日期：2024年12月16日；发布日期：2024年12月24日

摘　要

本文采用MATLAB软件，对河北省1978~2022年国内生产总值(GDP)进行分析，选取1978~2018年数据作为训练集，选取2019~2022年数据作为测试集，最终创建了ARIMA(1, 2, 0)模型，并使用此模型预测了河北省未来四年即2023~2026年的地区生产总值。根据此模型，预测出河北省2023~2026年GDP分别为：48679.4亿元、52075.9亿元、55132.5亿元和58518.9亿元。最终计算平均预测误差为1.91%，发现该模型具有较好的预测精度，能够有效的预测河北省GDP。

关键词

GDP，ARIMA模型，平均预测误差

Research on GDP Prediction of Hebei Province Based on ARIMA Model

Tao Ma[1], Jian Yu[1], Zexiang Yu[2], Hongxing Liu[1]

[1]College of Mathematics and Systems Science, Shenyang Normal University, Shenyang Liaoning
[2]Sydney Smart Technology College, Northeast University, Shenyang Liaoning

Received: Nov. 20th, 2024; accepted: Dec. 16th, 2024; published: Dec. 24th, 2024

Abstract

This article uses MATLAB software to analyze the Gross Domestic Product (GDP) of Hebei Province from 1978 to 2022. The data from 1978 to 2018 is selected as the training set, and the data from 2019 to 2022 is selected as the test set. Finally, an ARIMA (1, 2, 0) model is created, and this model is used to predict the regional GDP of Hebei Province in the next four years, namely 2023~2026. According to this model, it is predicted that the GDP of Hebei Province from 2023 to 2026 will be 48679.4 billion yuan, 52075.9 billion yuan, 55132.5 billion yuan, and 5851.89 billion yuan, respec-

文章引用: 马涛, 于涧, 于泽翔, 刘红星. 基于 ARIMA 模型的河北省 GDP 预测研究[J]. 统计学与应用, 2024, 13(6): 2375-2381. DOI: 10.12677/sa.2024.136230

tively. The final average prediction error is 1.91%, indicating that the model has good prediction accuracy and can effectively predict the GDP of Hebei Province.

Keywords

GDP, ARIMA Model, Average Prediction Error

1. 引言

GDP 可以反映一个国家或地区的经济发展规模，判断其经济总体实力和经济发展的快慢，可用来进行经济结构分析，如产业结构、需求结构和地区结构分析，是进行宏观经济决策的重要依据[1]。基于时间序列模型的 GDP 预测方法，由于其简单可行、高精度的优点，受到了人们的广泛关注。陈满丽[2]等运用 MATLAB 建立 ARIMA 模型，对我国 1978~2023 年国内生产总值进行分析，通过数据平稳性检验、模型检验等环节确定了 ARIMA (4, 2, 2)模型。殷佳棋[3]基于 SPSS 建立 ARIMA 模型，对内蒙古自治区 1990~2022 年的 GDP 数据进行分析，预测 2023~2027 年内蒙古自治区 GDP 将持续稳定增长。肖丹[4]运用 Python 软件对四川省 1978~2022 年的国内生产总值数据进行分析检验，最终创建了 ARIMA (2, 1, 0)模型，并使用此模型预测了四川省未来五年即 2023~2027 年的地区生产总值。

本文利用 MATLAB 软件进行编程建模，利用 ARIMA (p, d, q)模型，首先对河北省 1978~2022 年的 GDP 时间序列数据分别建立训练集和测试集，训练集用于模型的建立，测试集用于模型的检验，再对时间序列数据进行平稳性检验，最终确定时间序列数据在经过二阶差分处理后达到平稳状态，即确定 $d=2$。之后通过分析二阶差分后的时间序列数据的自相关图和偏自相关图的情况后，得到 p 和 q 的值分别为 1 和 0，确定模型为 ARIMA (1, 2, 0)，经训练集的检验后，此模型的平均预测误差为 1.91%，误差达到预期，模型通过检验。最终利用此模型预测出了河北省 2023~2026 年的 GDP 数据，可以为河北省未来的地区经济发展和宏观经济决策提供依据。

本文在第 2 节给出了 ARIMA 模型的理论依据和计算流程。第 3 节构建出了基于 ARIMA (1, 2, 0)的河北省 GDP 预测模型，并通过数据检验确定了其准确性。第 4 节对本文进行了总结与展望。

2. ARIMA 模型

2.1. 模型介绍

自回归差分移动平均模型(ARIMA 模型)主要包含自回归模型(AR)、差分过程(I)以及移动平均模型(MA)三部分。

AR 模型，即自回归模型，其优势是对于具有较长历史趋势的数据，AR 模型可以捕获这些趋势，并据此进行预测。但是 AR 模型不能很好地处理某些类型的时间序列数据，例如那些有临时、突发的变化或者噪声较大的数据。AR 模型相信“历史决定未来”，因此很大程度上忽略了现实情况的复杂性、也忽略了真正影响标签的因子带来的不可预料的影响[5]。

相反地，MA 模型，即移动平均模型，可以更好地处理那些有临时、突发的变化或者噪声较大的时间序列数据。但是对于具有较长历史趋势的数据，MA 模型可能无法像 AR 模型那样捕捉到这些趋势。MA

模型相信“时间序列是相对稳定的，时间序列的波动是由偶然因素影响决定的”，但现实中的时间序列很难一直维持“稳定”这一假设。

差分过程，用于使非平稳时间序列达到平稳，通过一阶或者二阶等差分处理，消除了时间序列中的趋势和季节性因素[6]。

基于以上模型的优缺点，引入了 ARIMA 模型，这是一种结合了 AR 模型和 MA 模型优点的模型。ARIMA 模型是一种在数据分析领域广泛应用且极为有效的方法。它以独特视角，试图借数据自相关性和差分挖掘深层信息。在时间序列数据里，自相关性宛如数据内的线索，ARIMA 模型抓住此线索并运用差分手段，巧妙提取出隐藏于数据背后、不易发现的时间序列模式，这些模式就像钥匙，可开启预测未来数据之门。它可以对数据中的趋势、季节性等复杂的特征进行准确剖析和捕捉，从而建立起科学合理的预测模型。通过这个模型，无论是经济数据的走势预测、气象数据的变化估计，还是其他众多领域中与时间序列相关的数据预测问题，都有了可靠的解决途径。

2.2. 模型计算流程

首先需要对时间序列数据进行平稳性检验，若通过平稳性检验，则可确定 d 值为 0。若未通过平稳性检验，对于非平稳时间序列要先进行 d 阶差分，转化为平稳序列。之后，需要确定 p 和 q 的值来建立模型，本文利用对时间序列数据的自相关图和偏自相关图的拖尾或截尾情况进行分析，得到最佳的阶数 p、q，由得到 p、q、d 的值，得到 ARIMA (p, d, q)模型，若模型通过检验，则可直接用来进行未来数据的预测。

3. 模型构建

3.1. 数据导入

文章数据来源于《河北统计年鉴 2023》，得到 1978~2022 年的国内生产总值 GDP，如表 1 所示。将数据分为训练组(1978~2018 年)和测试组(2019~2022 年)。其中训练组用于建立模型，测试组用于检验模型预测效果。

Table 1. Statistics of China’s gross domestic product from 1978 to 2022
表 1. 1978~2022 年中国国内生产总值统计

年份	GDP	年份	GDP	年份	GDP
1978	183.1	1993	1620.8	2008	14200.1
1979	203.2	1994	2114.5	2009	15306.9
1980	219.2	1995	2701.2	2010	18003.6
1981	222.5	1996	3198.0	2011	21384.7
1982	251.5	1997	3652.1	2012	23077.5
1983	283.2	1998	3924.5	2013	24259.6
1984	332.2	1999	4158.9	2014	25208.9
1985	396.8	2000	4628.2	2015	26398.4
1986	436.7	2001	5062.9	2016	28474.1
1987	521.9	2002	5518.9	2017	30640.8
1988	701.3	2003	6333.6	2018	32494.6

续表

1989	822.8	2004	7588.6	2019	34978.6
1990	896.3	2005	8773.4	2020	36013.8
1991	1072.1	2006	10043.0	2021	40391.3
1992	1278.5	2007	12152.9	2022	42370.4

注：单位：亿元。

利用 readtable 函数导入数据，并采用 plot 函数绘制时间序列图，如图 1 所示。

MATLAB 语法：

clc, clear all, close all;

Readtable (‘河北省地区生产总值(1978~2022 年).xlsx’)

time = ans. Time;

GDP = ans. GDP;

Plot (time, GDP)

[t, pValue, stat, cValue] = daftest (GDP)

语法中，若 t = 1，说明序列平稳；若 t = 0，说明序列不平稳，需对数据进行平稳化处理。若 pValue < 0.05，时间序列平稳。若 stat < cValue，时间序列平稳。反之，时间序列不平稳。

输出结果为：t = 0；pValue = 0.990；stat = 11.8503；cValue = −1.9472。

通过图 1 的折线图可以看出，1978~2022 年间，河北省 GDP 数据序列存在比较明显的长期递增形式，所以序列是属于非平稳的，因此我们要对 1978~2022 年的 GDP 数据进行平稳化处理。

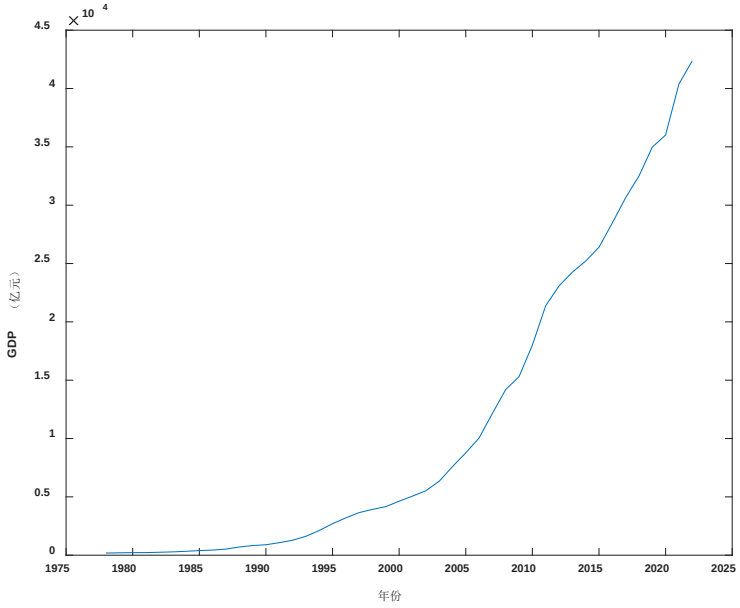

Figure 1. Time series chart of GDP in Hebei Province
图 1. 河北省 GDP 时序图

3.2. 数据平稳化处理

对数据进行一阶差分处理，得到一阶差分时序图如图 2 所示。

MATLAB 语句如下[7]：

GDPd1 = diff (GDP, 1)；

Plot (time (2: end), GDPd1)

[t1, pValue1, stat1, cValue1] = adftest (GDPd1)

输出结果为：$t_1 = 0$；$pValue_1 = 0.0810$；$stat_1 = -1.7176$；$cValue_1 = -1.9473$

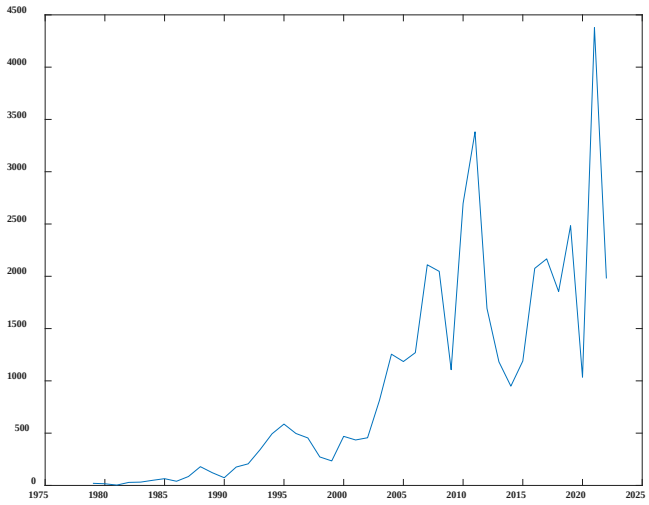

Figure 2. First-order differential timing diagram
图 2. 一阶差分时序图

通过图 2 的可以看出，数据序列仍然存在比较明显的递增形式，所以序列仍是属于非平稳的，因此需要继续进行二阶差分处理。

输出结果为：$t_2 = 1$；$pValue_2 = 1.0000\ e{-}03$；$stat_2 = -11.0856$；$cValue_2 = -1.9474$。

得到二阶差分时序图如图 3 所示。

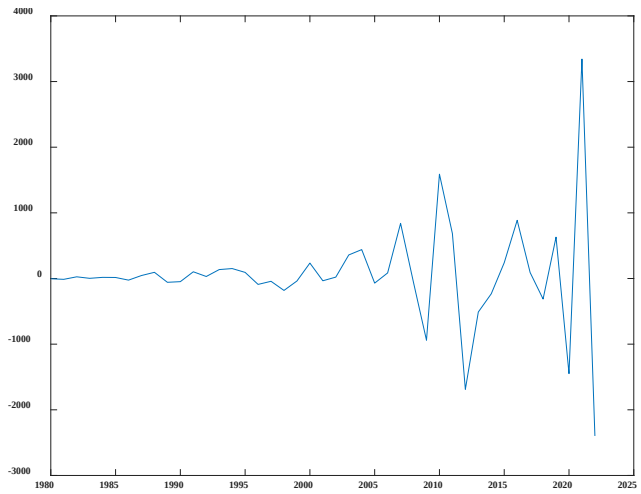

Figure 3. Second-order differential timing diagram
图 3. 二阶差分时序图

从图 3 的可以看出，数据在 0 附近波动，表现为平稳序列，通过平稳性检验。

3.3. 模型的建立及检验

ARIMA 模型参数主要有 p、d、q 三个，其中，参数 d 的估计值就是差分的阶数。因此，这里取 d = 2。对于参数 p，q，可以利用差分后平稳序列的自相关函数(ACF)和偏自相关函数(PACF)的统计特性选择合适的阶数，得到自相关函数图象呈一阶拖尾，偏自相关系数图象呈现截尾状态，因此选择 p = 1，q = 0，最终选择 ARIMA (1, 2, 0)对序列进行建模分析。

残差序列的随机性可以通过自相关函数法来检验，即做残差的自相关函数图，如图 4 所示。

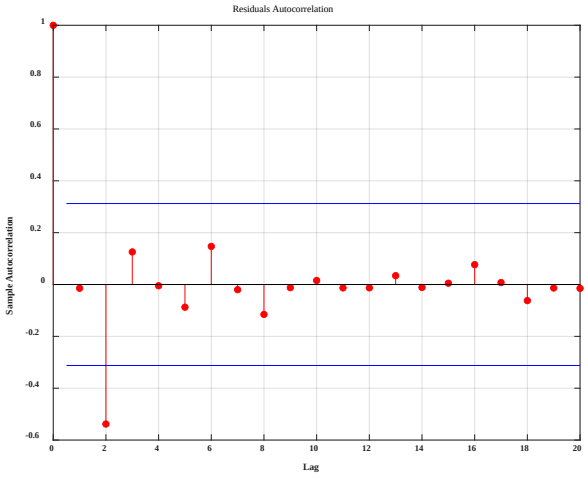

Figure 4. Autocorrelation of residuals
图 4. 残差的自相关性

从图 4 中可以看出各个残差之间的独立性比较高，因此模型通过检验。

3.4. 模型的预测及分析

通过建立的 ARIMA 模型，来进行时间序列的相关预测，利用函数 forecast [8]对 2019~2022 年的 GDP 进行预测。如表 2 所示，预测值的趋势与实际值的趋势保持一致，ARIMA (1, 2, 0)模型平均预测误差为 1.91%，预测精度较高。利用此模型，对河北省 2023~2026 年的 GDP 进行预测，得到 2023 年 GDP 为 48679.4 亿元，2024 年 GDP 为 52075.9 亿元，2025 年 GDP 为 55132.5 亿元，2026 年 GDP 为 58518.9 亿元。

Table 2. Statistics of China’s gross domestic product from 1978 to 2022
表 2. 1978~2022 年中国国内生产总值统计

年份	实际值	预测值	预测误差	平均预测误差
2019	34978.6	34405.3	1.64%	1.91%
2020	36013.8	36363	0.97%	
2021	40391.3	39368	2.53%	
2022	42370.4	43420.4	2.48%	
2023	未知	48679.4	未知	未知
2024	未知	52075.9	未知	未知
2025	未知	55132.5	未知	未知
2026	未知	58518.9	未知	未知

注：单位：亿元。

4. 总结与展望

在本文的工作中，我们基于河北省 1978~2022 年的 GDP 数据构建模型，在文章的第一部分中介绍了该问题模型的研究背景、意义以及前人所做工作。在第二部分中，我们对模型的相关参数、概念和计算流程进行了简要的介绍，为接下来的模型构建奠定了基础。在第三部分中，我们进行模型的构建，首先对数据进行平稳性检验，发现数据在进行二阶差分后表现为平稳序列，之后分析自相关函数和偏自相关函数图像的拖尾、结尾情况确定 p、q 的值，最终确定模型为 ARIMA (1, 2, 0)。模型通过残差平稳性检验，证明该模型可以很好的预测河北省未来 GDP 数据。但该模型的预测精度仍然存在较大误差，并且对未来数据的预测精准度也需要后续的验证和比对。对于模型预测精度的提升问题以及该模型能否与其他模型进行综合预测等问题，未来将成为我们主要的研究方向。

基金项目

教育部就业育人项目(2024011234627)：基于物联网的大数据就业趋势研究。

参考文献

[1] 于涧, 马涛, 于泽翔, 等. 基于 BP 神经网络的河北省 GDP 预测研究[J]. 北方工业大学学报, 2024, 36(3): 126-130.

[2] 陈满丽, 张慧娟, 焦楠楠, 等. 基于 MATLAB 的 ARIMA 模型对我国 GDP 预测的研究[J]. 中国市场, 2024(12): 1-4.

[3] 殷佳棋. 内蒙古自治区 GDP 分析与预测[J]. 内蒙古科技与经济, 2024(7): 15-18, 54.

[4] 肖丹. 基于 ARIMA 模型的四川省 GDP 分析与预测[J]. 生产力研究, 2023(10): 62-66.

[5] 刘泽华, 卢洪涛, 李伟, 等. 基于 R 语言的 ARIMA 模型在医用耗材消耗量短期预测中的应用研究[J]. 医疗卫生装备, 2024, 45(10): 84-87.

[6] 詹平, 刘飞翔, 赵嘉良. 基于 LDA 和 ARIMA 模型的煤矿安全隐患数量预测研究[J]. 煤, 2024, 33(3): 39-44.

[7] 吴会会, 王嘉鹏, 吴文静, 等. 基于 ARIMA 模型的全球气表温度预测分析[J]. 现代信息科技, 2023, 7(16): 147-150.

[8] Gao, W., Xiao, T., Zou, L., Li, H. and Gu, S. (2024) Analysis and Prediction of Atmospheric Environmental Quality Based on the Autoregressive Integrated Moving Average Model (ARIMA Model) in Hunan Province, China. *Sustainability*, **16**, Article 8471. https://doi.org/10.3390/su16198471