

基于ARIMA模型的地区国内旅游人数研究 ——以云南省为例

周子渊

云南省文化和旅游信息中心数据业务科, 云南 昆明

收稿日期: 2024年6月1日; 录用日期: 2024年6月21日; 发布日期: 2024年6月30日

摘要

云南旅游资源丰富且得天独厚, 经过多年的发展云南旅游业已成为国民经济中的支柱产业, 作为国内主要目的地, 2023年, 云南国内旅游人数已突破10亿人次, 当前正值启动“十五五”规划的起步阶段, 如何科学预测云南国内旅游人数趋势, 是一项重要课题。本文通过对2010~2023年云南省国内游客季度数据的研究, 运用时间序列模型中ARIMA模型进行拟合, 选取拟合模型中最优的模型ARIMA (2, 1, 3), 得到拟合函数, 并预测未来3年云南省国内旅游人数季度趋势, 以此给予政府在接下来经济规划中提供相关建议。

关键词

国内旅游人数, ARIMA, 季节趋势模型, 云南省

Research on the Number of Domestic Tourists in a Region Based on ARIMA Model —Taking Yunnan Province as an Example

Ziyuan Zhou

Department of Data Business, Yunnan Provincial Culture and Tourism Information Center, Kunming Yunnan

Received: Jun. 1st, 2024; accepted: Jun. 21st, 2024; published: Jun. 30th, 2024

Abstract

Yunnan has abundant and unique tourism resources. After years of development, Yunnan tourism has become a pillar industry in the national economy. As a major destination in China, the number of domestic tourists in Yunnan has exceeded 1 billion in 2023. Currently, it is in the initial stage of

launching the “15th Five Year Plan”. How to scientifically predict the trend of domestic tourists in Yunnan is an important issue. This article studies the quarterly data of domestic tourists in Yunnan Province from 2010 to 2023, uses the ARIMA model in the time series model for fitting, selects the best model in the fitting model ARIMA (2, 1, 3), obtains the fitting function, and predicts the quarterly trend of domestic tourists in Yunnan Province in the next three years, in order to provide relevant suggestions for the government in future economic planning.

Keywords

Number of Domestic Tourists, ARIMA, Seasonal Trend Model, Yunnan Province

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

1.1. 背景

云南旅游资源丰富, 优良多样的生态环境、神奇美丽的自然风光、多姿多彩的民族风情、源远流长的历史文化、多元包容的宗教信仰和得天独厚的区位优势, 为云南旅游产业的发展提供了良好的条件。改革开放以来, 历届省委、省政府都高度重视旅游产业发展, 依托得天独厚的资源优势, 顺应国际旅游发展趋势, 实施政府主导型发展战略, 经历了起步发展(1978 至 1988 年间主要以单纯的事业型接待服务为主)、产业建设(1989 至 1995 年间从“接待事业型”向“经济产业型”转变)、支柱产业建设(1996 至 2005 年间成为支柱产业)、“二次创业”(2005 至 2013 年间实施优化结构、转型升级、提质增效的“旅游二次创业”)、旅游强省建设(2013 年 9 月到至今)等阶段, 实现了旅游产业从无到有、从小到大的历史性飞跃, 走出了一条符合自身实际的特色旅游发展路子。当前, 云南旅游业经历了高速发展后, 正处于一个稳定发展的阶段, 所以在未来旅游经济规划上, 一个好的风向标将带领云南省国内旅游人数走出一个新的高度。本文就云南省 2010 年至 2023 年的季度国内旅游人数进行了分析, 拟合合适的模型以估计后三年云南省季度国内旅游人数, 从结果中为旅游主管部门制定旅游业发展和“十五五”文旅规划提供部分参考。

1.2. 相关文献综述

在有关国内旅游人数方面的研究中, 比较偏少, 多数国内学者主要侧重对地区生产总值的研究, 并建立模型进行预测研究。按照研究方法可以大致分为计量经济法(回归模型等)、时间序列法(ARIMA 模型、移动平均法、ARCH 模型、误差修正模型等)、人工智能法(主要包含有神经网络、灰色理论方法等)。

在时间序列的相关预测方法中, 刘金娣, 卢婷艳[1]曾运用误差修正模型分析地区生产总值与政府消费之间存在长期均衡的协整关系, 经过分析得到政府消费的当期波动以及上期误差(ECM)对地区生产总值的当期波动有显著影响, 由此得到提出建议, 提高政府消费是实现经济增长的一个有效途径。刘展, 赵明霞, 郭丽娟[2]曾基于时间序列方法中的求和自回归移动平均模型(ARIMA 模型)对平顶山地区生产总值的变化趋势进行了相应的预测, 并经过分析比较得出了时间序列预测法在平顶山地区生产总值预测中的应用具有良好效果的结论, 以此为基础, 为平顶山各处生产总值的变化趋势建立相应的 ARIMA 模型。宋静[3]利用时间序列理论, 在检验云南省国内旅游人数的平稳性时, 发现数据非平稳, 于是对其进行平稳化处理, 再对数据进行 AIC 定则检验, 最终确定最理想的模型为 ARIMA(1,1,2), 在预测值与真实值之

间的误差分析后认为该模型可行性高，并且在 2020 年预测之后发现超过了 2020 年的目标 GDP。

在神经网络方法中，李云晶，廖飞[4]利用 GM (1,1) 模型，预测 2020~2025 年黑龙江省的人均地区生产总值，在研究中发现，黑龙江省人均地区生产总值保持在未来的几年着相对稳定的增长趋势。卢俊岚，王明辉[5]根据灰色预测法的相关理论，研究了广东省 1978~2016 年的地区生产总值，最后根据数据拟合结果，预测广东省未来几年经济发展水平。姚敏奇[6]通过对 R 语言知识的运用，利用 GM (1,1) 灰色预测模型研究石家庄市区域经济增长的相关问题。根据“十二五”期间石家庄市的相关数据建立 GM (1,1) 模型，并对模型进行检验，并且通过模型对石家庄市经济增长情况进行预测以及分析。

在回归模型中，彭思嘉[7]通过广义差分法对模型进行回归，在对回归方程先后进行 ADF 检验和 Granger 因果检验，由检验结果得知，湖南省的财政总支出和生产总值变量都是非平稳时间序列，但是通过对其进行单位根的二阶差分序列检验之后，发现差分后的线性组合是平稳序列，由此可见二者之间存在着一定的稳定均衡的关系，在此结论下给出相关建议。刘旦男[8]利用逐步回归法和 ARCH 模型，分析了影响湖南省地区生产总值的因素，结合指数平滑法，对湖南 2019 年至 2020 年的地区生产总值进行预测，通过结果可以看出对湖南地区生产总值影响的相关因素并提出合理建议。刘丽华，刘尧[9]通过对各地区 2007~2011 年地区生产总值的横向与纵向比较统计图分析，找出本年度生产总值最高的地区(江苏省)再根据江苏省地区生产总值数据，借用多元线性回归模型来研究影响江苏省生产总值的因素。

通过上面的文献分析，以及拟使用的 2010 年~2023 年云南省国内旅游人数季度数据，根据选用的数据特征，本文将选用时间序列模型进行数据建模，但是具体采用什么模型则根据各项检验和估计来进行识别，最后对将来 3 年云南省国内旅游人数进行预测。

2. 模型介绍

2.1. 纯随机性介绍

如果时间序列 $\{x_t\}$ 满足如下条件：

任意取 $t \in T$ ，有 $Ex_t = \mu$ ；

任意取 $t, s \in T$ 有 $\gamma(t, s) = \begin{cases} \sigma^2, & t = s \\ 0, & t \neq s \end{cases}$

则可以称序列 $\{x_t\}$ 是纯随机序列，简记为 $x_t \sim WN(\mu, \sigma^2)$ 。

白噪声具有的性质如下：

$$\gamma(k) = 0, \forall k \neq 0$$

说明白噪声的各项之间没有任何相关关系，若序列 $\{x_t\}$ 为白噪声，则说明序列 $\{x_t\}$ 中没有值得提取的有用的信息，于是我们应该终止该序列的后续分析。

2.2. ARIMA 模型介绍

在对时间序列平稳性进行检验时，时间序列非平稳，则可对原序列进行差分提取确定性信息后再建立模型。此时可以对原序列建立 ARIMA(p, d, q) 模型，模型中的 p 为自回归阶数， d 为消除非平稳序列需要差分的次数， q 为移动平均阶数。具体结构如下：

$$\begin{cases} \Phi(B)\nabla^d X_t = \Theta(B)\varepsilon_t \\ E(\varepsilon_t) = 0, \text{Var}(\varepsilon_t) = \varepsilon_t^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ E(x_t \varepsilon_s) = 0, \forall s < t \end{cases} \quad (1)$$

式中 $\nabla^d = (1-B)^d$, $\Phi(B) = 1 - \varphi_1 B - \dots - \varphi_p B^p$, 为平稳可逆自回归移动平均模型的自回归系数多项式;
 $\Theta(B) = 1 - \theta_1 B - \dots - \theta_q B^q$, 为自回归移动平均模型的平滑系数多项式。

2.3. 季节模型介绍

时间序列的波动一般分为四类要素：其中 T 为长期趋势要素、S 为季节要素、C 为经济周期循环要素、I 为以及不规则要素。常用季节调整分解形式有加法模型和乘法模型两种：

简单季节模型：

$$Y_t = TC_t + S_t + I_t \quad (1)$$

乘积季节模型：

$$Y_t = TC_t \times S_t \times I_t \quad (2)$$

式中, TC_t 表示的是趋势循环, 这是因为在调整季节因素的时候并不能把长期趋势和循环要素分开。

简单季节模型各种效应信息在提取的时候非常容易, 为了将序列中的季节信息提取充分, 我们仅仅通过简单的周期步长差分即可完成, 同时, 通过简单的低阶差分即可将趋势信息提取充分, 在提取完季节信息和趋势信息之后, 我们所得到的残差序列就是一个平稳序列。其模型结构如下：

$$\nabla_D \nabla^d x_t = \frac{\Theta(B)}{\Phi(B)} \varepsilon_t \quad (3)$$

其中：

D 为周期步长, d 为提取趋势信息用的差分阶数；

$\{\varepsilon_t\}$ 为白噪声序列, 且 $E(\varepsilon_t) = 0, \text{var}(\varepsilon_t) = \sigma_\varepsilon^2$ ；

$\Theta(B) = 1 - \theta_1 B - \dots - \theta_q B^q$ 为 q 阶移动平均系数多项式；

$\Phi(B) = 1 - \varphi_1 B - \dots - \varphi_p B^p$ 为 p 阶自回归系数多项式。

当序列存在季节效应并且季节效应本身存在相关性时, 可以使用以周期步长为单位的 $ARMA(P, Q)$ 模型进行提取。又因为短期相关性和季节效应之间存在乘积关系, 所以出来的拟合模型为 $ARMA(P, Q)$ 和 $ARMA(p, q)$ 的乘积。乘积模型的表达式如下：

$$\nabla^d \nabla_s^D x_t = \frac{\Theta(B)\Theta_s(B)}{\Phi(B)\Phi_s(B)} \varepsilon_t \quad (4)$$

其中：

$$\Theta(B) = 1 - \theta_1 B - \dots - \theta_q B^q$$

$$\Phi(B) = 1 - \varphi_1 B - \dots - \varphi_p B^p$$

$$\Theta_s(B) = 1 - \theta_1 B^S - \dots - \theta_Q B^{QS}$$

$$\Phi_s(B) = 1 - \varphi_1 B^S - \dots - \varphi_P B^{PS}$$

2.4. ADF 检验

为了避免在研究时间序列的过程中出现虚假回归的问题, 所以在拟合动态回归模型时, 必须先对序列的平稳性进行检验, 只有序列平稳才可以拟合时间序列的动态回归关系。

ADF 检验适用于 $AR(p)$ 模型的平稳性检验, 其原理如下：

对任意 $AR(p)$ 模型

$$x_t = \varphi_1 x_{t-1} + \cdots + \varphi_p x_{t-p} + \varepsilon_t \quad (5)$$

其特征方程为:

$$\lambda^p - \varphi_1 \lambda^{p-1} - \cdots - \varphi_p = 0 \quad (6)$$

如果方程所有的特征根都在单位圆内, 即

$$|\lambda_i| < 1, i = 1, 2, 3, p$$

则序列 $\{x_t\}$ 平稳。

如果方程有一个单位根, 假设

$$\lambda_1 = 1$$

则序列 $\{x_t\}$ 非平稳, 并且自回归系数之和恰好等于 1:

$$\lambda^p - \varphi_1 \lambda^{p-1} - \cdots - \varphi_p = 0 \Rightarrow \varphi_1 + \varphi_2 + \cdots + \varphi_p = 1$$

所以, 检验得到自回归系数和为 1 时, 就可以判定序列非平稳。

对公式(5)等价变换, 得到:

$$\begin{aligned} x_t - x_{t-1} &= \varphi_1 x_{t-1} + \cdots + \varphi_p x_{t-p} - x_{t-1} + \varepsilon_t \\ &= (\varphi_2 + \cdots + \varphi_p) x_{t-1} + \varphi_1 x_{t-1} - x_{t-1} - (\varphi_2 + \cdots + \varphi_p) x_{t-1} \\ &\quad + \varphi_2 x_{t-2} + (\varphi_3 + \cdots + \varphi_p) x_{t-2} - (\varphi_3 + \cdots + \varphi_p) x_{t-2} \\ &\quad + \varphi_3 x_{t-3} + (\varphi_4 + \cdots + \varphi_p) x_{t-3} + \cdots - \varphi_p x_{t-p+1} + \varphi_p x_{t-p} + \varepsilon_t \end{aligned}$$

整理以上公式得到:

$$\nabla x_t = (\varphi_1 + \cdots + \varphi_p - 1) x_{t-1} - (\varphi_2 + \cdots + \varphi_p) \nabla x_{t-1} - \cdots - \varphi_p \nabla x_{t-p+1} + \varepsilon_t$$

简单记做:

$$\nabla x_t = \rho x_{t-1} + \beta_1 \nabla x_{t-1} + \cdots + \beta_{p-1} \nabla x_{t-p+1} + \varepsilon_t \quad (7)$$

其中:

$$\rho = \varphi_1 + \cdots + \varphi_p - 1$$

$$\beta_j = -(\varphi_{j+1} + \cdots + \varphi_p), j = 1, 2, 3, p-1$$

所以, 若序列 $\{x_t\}$ 平稳, 则:

$$\varphi_1 + \cdots + \varphi_p < 1 \Leftrightarrow \rho < 0$$

若序列 $\{x_t\}$ 非平稳, 则:

$$\varphi_1 + \cdots + \varphi_p = 1 \Leftrightarrow \rho = 0$$

所以在 AR(p)单位根检验过程中, 假设检验可以确定为如下形式:

$$H_0: \rho = 0 (\text{序列}\{x_t\} \text{非平稳}) \leftrightarrow H_1: \rho < 0 (\text{序列}\{x_t\} \text{平稳})$$

构造的 ADF 统计量:

$$\tau = \frac{\hat{\rho}}{S(\hat{\rho})}$$

其中, $S(\hat{\rho})$ 为参数 ρ 的样本标准差。

3. 模型的建立

3.1. 数据来源

本文选择的数据为 2010 年~2023 年的云南省国内旅游人数季度数据，数据均来源于云南省文化和旅游厅。从获得的数据来看该数据每年从第一季度到第四季度都有一个增长的趋势，所以初步认为该数据具有明显的季节趋势，适合使用季节模型进行拟合。

3.2. 平稳性检验

时间序列分析要求数据平稳且非白噪声，所以导入数据之后第一步就是检查数据的平稳性。将转化为时间序列后的数据命名为 x ，绘制时序图(见图 1)，图一显示云南省国内旅游人数季度数据序列具有趋势和周期，绘制自相关图(见图 2)，更加确定序列明显不平稳，需要对数据一阶差分之后才能建模。

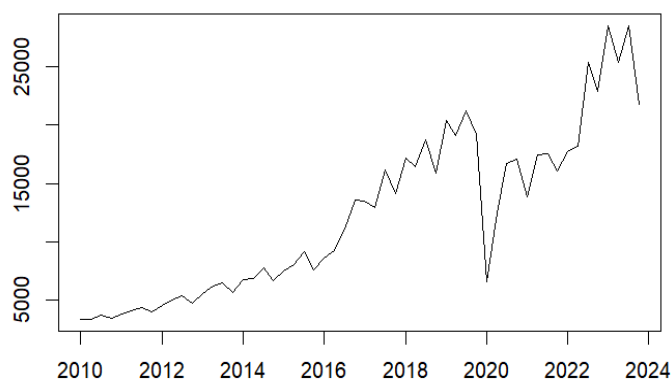


Figure 1. Time series diagram

图 1. 时序图

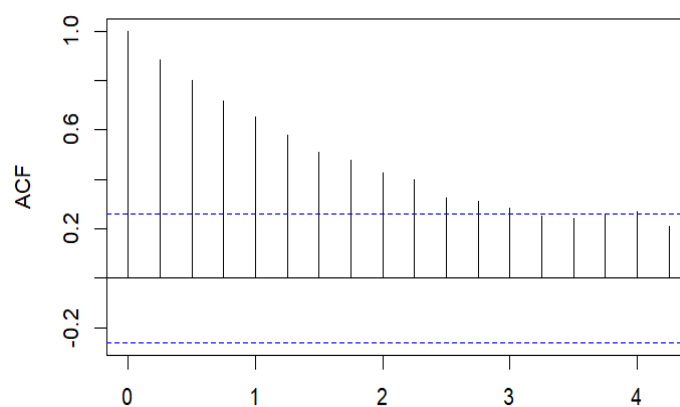


Figure 2. Autocorrelation graph

图 2. 自相关图

3.3. 非平稳数据的处理

由图 1 可以看出来云南省国内旅游人数具有一个线性增长趋势，结合图 2，为了提取原序列的趋势信息我们可以对其进行 1 阶差分，想要提取周期信息则可以通过四步差分进行，在对原数据进行 1 阶 4 步差分后得到的时序图所示(见图 3)。

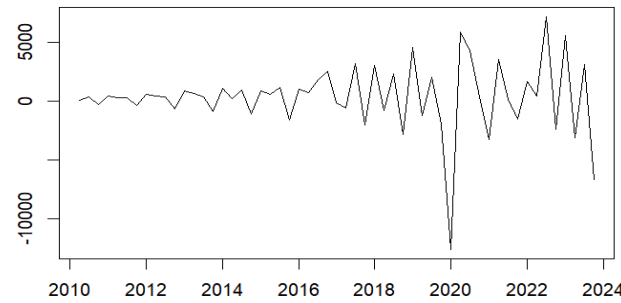


Figure 3. 1st order differential timing diagram
图 3. 1 阶差分时序图

观察 1 阶差分后的序列，已经没有明显的趋势，同时发现季节性影响已经不明显。

对序列进行平稳性检验及白噪声检验，平稳性检验 ADF 统计量的 P 值为 0.01，小于 0.05，可以认为差分后的序列是平稳的；

对差分后的序列进行滞后 4 期的白噪声检验，残差白噪声检验 p 值分别为 0.008255, 0.02703, 0.02114, 0.01941，由于残差白噪声检验 p 值均小于 0.05，由此我们可以得到结论，差分后的序列属于非白噪声，我们可以进行接下来的分析。

3.4. 模型的建立与检验

对差分后的序列进行 acf 检验和 pacf 检验，得到如下图所示(见图 4)：

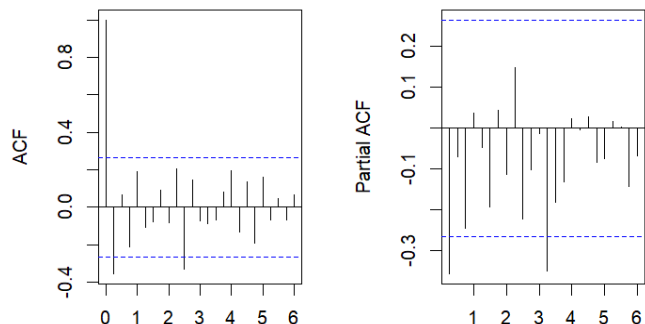


Figure 4. Autocorrelation and partial autocorrelation graphs
图 4. 自相关图和偏自相关图

根据差分序列的自相关图和偏自相关图中的滞后阶数，可以拟定以下模型(见表 1)。

Table 1. Comparison of models
表 1. 模型的比较

模型	AIC 值	1 阶白噪声检验 P	2 阶白噪声检验 P
模型一： $ARIMA(1,1,0) \times (1,0,0)_4$	1033	0.7812	0.7418
模型二： $ARIMA(0,1,1) \times (1,0,0)_4$	1032.76	0.761	0.9536
模型三： $ARIMA(0,1,1) \times (1,0,1)_4$	1034.75	0.7648	0.9557
模型四： $ARIMA(1,1,2)$	1035.54	0.8204	0.6222
模型五： $ARIMA(0,1,1)$	1032.46	0.5841	0.8591

五个模型中,残差白噪声检验 P 值均大于 0.05,说明五个模型均能提取序列中的有效信息,但根据 AIC 值可以发现模型三更优,故 $ARIMA(2,1,3)$ 为本文根据数据特征建立的最优云南省国内旅游人数预测模型,其具体形式为:

$$x_t = \frac{(1+1.0283B+0.1943B^2-0.3904B^3)}{1-1.3106B-0.5644B^2} \varepsilon_t, \quad \varepsilon_t \sim N(0.6683827)$$

对于检验结果,首先观察 $ARIMA(2,1,3)$ 的残差自相关图(见图 6),发现残差随机性非常强,对模型进行 1 阶和 2 阶残差白噪声检验,结果如表 2,白噪声检验表明此残差序列可以视为白噪声,说明该拟合模型可以作为云南省旅游人数季度数据的拟合模型。

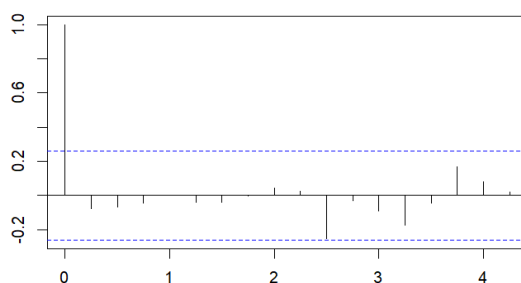


Figure 5. Residual autocorrelation graph

图 5. 残差自相关图

3.5. 模型的预测

此时 $ARIMA(2,1,3)$ 模型可以用来预测,未来 3 年的得到的云南省季度国内旅游人数如下表所示(见表 2),预测图如下图所示(见图 6)。

Table 2. Model prediction

表 2. 模型的预测

季度	2024 年	2025 年	2026 年
I	25067.08	24707.12	24045.49
II	22993.67	24154.19	24416.18
III	25458.09	24140.23	24170.24
IV	23398.29	24470.57	24283.38

2024 年 1 季度云南省国内旅游人数实际 3.0 亿人次,在 95%置信区间(19593.07, 30300.21)内。

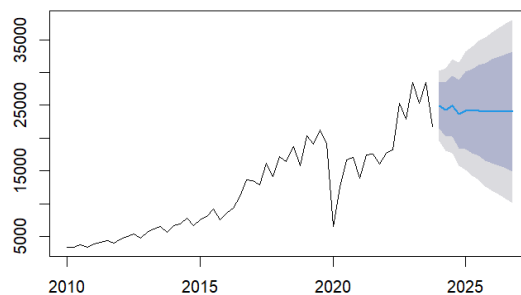


Figure 6. Prediction chart

图 6. 预测图

4. 结论及建议

本文基于 2010 年至 2023 年云南省国内旅游人数数据,建立了模型,根据分析发现,云南省国内旅游人数受季节性影响变化不明显,最后选择了最优模型 $ARIMA(2,1,3)$ 对云南省 2024 年至 2026 年中每个季度做了预测,预测结果相比刚过去的 13 年,云南省季度国内旅游人数总体呈现上升趋势且季度具有明显的周期性。

通过数据预测可以了解云南省未来三年国内旅游人数发展走向,通过图 6 可以看出未来几年云南省国内旅游人数处于一个较快的发展阶段,进一步说明云南省政府采取的经济策略是有效的正确的,尤其是在第四季度,国内旅游人数及其发展速度显著高于第一第二季度,所以云南省政府在制定接下来第一季度和第二季度发展战略时,可以借鉴第四季度,将云南省国内旅游人数进一步提高。

参考文献

- [1] 刘金娣, 卢婷艳. 关于山东省政府消费与地区生产总值的研究——基于协整检验与误差修正模型[J]. 现代营销, 2017(5): 210.
- [2] 刘展, 赵明霞, 郭丽娟. 基于时间序列分析的平顶山地区生产总值的变化趋势研究[J]. 河南科学, 2013, 31(10): 1739-1744.
- [3] 宋静. 时间序列在安徽省 GDP 预测中的应用——基于 ARIMA 模型[J]. 商场现代化, 2016(21): 134-136.
- [4] 李云晶, 廖飞. 灰色预测模型在黑龙江省人均地区生产总值预测中的应用[J]. 产业创新研究, 2020(12): 44-45.
- [5] 卢俊岚, 王明辉. 基于灰色预测法对广东省地区生产总值的预测分析[J]. 高师理科学刊, 2019, 39(1): 10-12, 17. <https://doi.org/10.3969/j.issn.1007-9831.2019.01.004>
- [6] 姚敏奇. 运用灰色关联分析预测十三五期间石家庄市地区生产总值[J]. 纳税, 2018(29): 2.
- [7] 彭思嘉. 地方政府支出与地区生产总值的相关性研究——以湖南省为例[J]. 中国市场, 2015(25): 115-116. <https://doi.org/10.13939/j.cnki.zgsc.2015.25.115>
- [8] 刘旦男. 基于计量模型对湖南地区生产总值影响因素分析[J]. 商讯: 商业经济文荟, 2019(6): 104-106, 146.
- [9] 刘丽华, 刘尧. 基于回归分析的地区生产总值变化研究[J]. 企业导报, 2013(9): 147-148.