Statistics and Application 统计学与应用, 2022, 11(5), 1132-1142
Published Online October 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.115116

Hans汉斯

ARIMA模型在陕西全社会固定资产投资预测中的应用

李　彤

云南财经大学统计与数学学院，云南　昆明

收稿日期：2022年9月18日；录用日期：2022年10月8日；发布日期：2022年10月20日

摘　要

本文从陕西省全社会固定资产投资情况出发，利用R软件分析了从国家统计局数据库搜集获得的陕西省1978至2020年陕西全社会固定资产投资情况，分析获得ARIMA(2,2,0)模型，并预测2021年之后陕西省的全社会固定资产投资额，建议政府调整经济发展模式，完善监管制度，结合中国社会现实情况，在社会健康发展的基础上进行社会经济建设，促进社会经济健康稳定的发展。

关键词

ARIMA模型，时间序列，预测

Application of ARIMA Model in the Prediction of Fixed Assets Investment in Shaanxi Province

Tong Li

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan

Received: Sep. 18th, 2022; accepted: Oct. 8th, 2022; published: Oct. 20th, 2022

Abstract

Starting from the fixed asset investment of the whole society in Shaanxi Province, this paper uses R software to analyze the fixed asset investment of the whole society in Shaanxi Province from 1978 to 2020 collected from the database of the National Bureau of statistics, and obtains the ARIMA(2,2,0) model, predicts the fixed asset investment of the whole society in Shaanxi Province

文章引用: 李彤. ARIMA 模型在陕西全社会固定资产投资预测中的应用[J]. 统计学与应用, 2022, 11(5): 1132-1142.
DOI: 10.12677/sa.2022.115116

after 2021, and suggests that the government adjust the economic development model, improve the regulatory system, and combine the reality of Chinese society, carry out social and economic construction on the basis of healthy social development, and promote healthy and stable social and economic development.

Keywords

ARIMA Model, Time Series, Forecast

1. 绪论

全社会固定资产投资总额是以货币表现的建造和购置固定资产活动的工作量，它是反映固定资产投资规模、速度、比例关系和使用方向的综合性指标。按照管理渠道，全社会固定资产投资总额分为基本建设、更新改造、房地产开发投资和其他固定资产投资四个部分。固定资产投资是社会固定资产再生产的主要手段，通过建造和购置固定资产的活动，可以不断采用先进技术装备，建立新兴部门，进一步调整经济结构和生产力的地区分布，可以直接促进经济增长，拉动内需，对于经济的持续、健康、稳定发展具有重要的意义。

改革开放以来，我国的经济发展迅速，不断地创下历史新高。我们知道，投资是拉动经济增长的三驾马车之一。1980 年仅为 910.9 亿元；2006 年则猛增至 109998.2 亿元。尤其是进入 21 世纪以来，外商投资的显著增加不仅推动了我国经济政策的调整与完善，而且也给经济增长增添了新的活力。全社会固定资产投资总额也持续增加对我国的经济持续增长有着非常重要的作用。对一个国家如此，对陕西省更是如此。陕西省处于我国的腹中之地，起着链接南北和东西部的重任。陕西省不临海，对外贸易虽逐年有递增的趋势，但是对外贸易量相对来说比较少。正因为地理位置的特殊性，导致成本较高，海外投资量也相对较少。所以，研究陕西省的经济的发展有着重要的意义。对于陕西省来说，我们需要了解在过去经济的发展中，陕西全社会固定资产投资对于陕西省经济的发展的影响有多大。对全社会固定资产投资有影响的因素很多，而这些因素彼此之间又有着错综复杂的联系。

ARIMA 模型要求时间序列数据平稳，并且只需要内生变量而不需要借助其他外生变量。因此，模型相较于其他模型比较简便。本文从动态角度考察，把我国全社会固定资产投资总额看成是一个时间序列，利用历史数据、运用统计学和计量经济学原理，从时间序列的定义出发，结合统计软件 R，运用 ARIMA 建模方法，将 ARIMA 模型应用于陕西历年全社会固定资产投资数据的分析与预测，给政府提出可行的建议，促进经济的健康发展。

2. 文献综述

前人对于全社会固定资产预测做了一些研究，主要是基于时间序列模型、灰色理论以及多因素相关模型对相关的时间序列进行预测，都得到了较好的结果。时间序列模型的应用可以很好地预测短期数据，对于序列的预测能够很好地为政府提供可靠的依据。而多因素相关模型从纵向和横向对各个模型进行了对比，为数据的预测提供了依据。对于灰色理论来说，他可以应用少量数据就能够对数据进行建模。总的来说这三种预测方法可以使用不同的场景，各有千秋。

2010 年李惠在文中，通过对 1980 年到 2007 年我国全社会固定资产投资的数据运用 ARIMA 方法建模，检验得出 ARIMA(4,2,4)模型最优[1]；靳宝琳和赫英迪用时间序列建模方法对太原市的固定资产投资总额进行了分析，建立了 ARIMA 模型。结果显示的预测效果较为准确，可用于未来的预测[2]；石美娟在文中采用自回归移动平均法，对固定资产投资进行了分析。也得到了比较准确的预测效果，可用于未来的预测[3]。蒋艳，庞林旗运用北京市 2000~2017 年来源法卫生总费用的相关数据，构建 Logistic 函数模型，对相关序列进行了预测研究。结论北京市卫生总费用增长较快，后期趋于平稳，应根据预测结果合理调控卫生总费用的发展[4] [5] [6]。

邓奎，李龙国应用数据加载法提出了 GM(1,1)的修正模型，通过灰色预测法和马尔柯夫链预测法的耦合，建立了城市污水排放量的灰色马尔柯夫预测模型.灰色马尔柯夫预测模型具有灰色系统应用少量数据即可建模，以及马尔柯夫链预测可以预测数据值波动较大的序列的特点.计算结果表明，城市污水排放量的预测值很好地吻合了实际值[7]。马占青，崔广柏，杨宏杰，王晓强等人应用数据加载法提出了 GM(1,1)的修正模型，通过灰色预测法和马尔柯夫链预测法的耦合，建立了城市污水排放量的灰色马尔柯夫预测模型。结果表明，城市污水排放量的预测值很好地吻合了实际值[8] [9]。

丘健明，王纯，李斌应用多因素相关模型，从纵向和横向两个方面对比各个模型，为数据的分析预测提供依据[10]，通过对多因素的分析，文章明确的分析对比了各因素对于固定资产投资的影响的程度，为相关部门更好的了解各因素对其的影响程度提供了有力的依据，以便在之后的政策制定的方向上提供好的向导作用。

3. ARIMA 模型简介

3.1. ARIMA 模型原理

首先可以将这个随机序列拟合成一个数学模型，并识别出该模型，然后使用该模型根据过去值与当前值预测未来值。

当时间序列出现长期趋势、随机扰动、季节性情况此时的时间序列并不稳定，假如存在非平稳的时间序列，若将其变得稳定，可以通过差分来实现这个要求。数学表达式如下：

$$\begin{cases}\Phi(B)\Delta^d x_t=\theta(B)\varepsilon_t\\ E(\varepsilon_t)=0, Var(\varepsilon_t)=\sigma_\varepsilon^2, E(\varepsilon_t\varepsilon_s)=0, s\neq t\\ E(\varepsilon_t\varepsilon_s)=0, \forall s<t\end{cases}$$

其中，$\Delta^d=(1-B)^d$；$\Phi(B)=1-\phi_1 B-\cdots-\phi_{11}B^p$；$\theta(B)=1-\theta_1 B-\cdots-\theta_1 B^p$。

如果时间序列有明显的趋势，可以通过一阶差分或二阶差分进行处理。时间序列是固定周期，一般以最小周期长度作为差值的步长。并且我们不能过度差分。如果差分阶数过多会导致，那么有效数据的信息将会受到很大影响，结果会导致数据信息丢失。

差分公式如下：

x_t 的一阶差分：$\nabla x_t=x_t-x_{t-1}$；

x_t 的二阶差分：$\nabla x_t^2=\nabla x_t-\nabla x_{t-1}$；

以此类推，x_t 的 d 阶差分：$\Delta^d x_t=\Delta^{p-1}x_t-\Delta^{p-1}x_{t-1}$。

3.2. 时间序列平稳性检验

如果时序图围绕着一个恒定的常数值上下进行波动而且波动幅度具有明显的特征则是平稳的。否则，就是非平稳序列。

3.2.1. ADF 检验原理

对任一 AR(p)过程：

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t$$

其特征方程为：

$$\lambda^p - \phi_1 \lambda^{p-1} - \cdots - \phi_p = 0$$

当$|\lambda_i| < 1, i = 1,2,\cdots,p$时，则该序列平稳。

或$\phi_1 + \phi_2 + \cdots + \phi_p = 1$

$$\nabla X_t = \rho X_{t-1} + \beta_1 \nabla X_{t-1} + \beta_2 \nabla X_{t-2} + \cdots + \beta_p \nabla X_{t-p} + w_t$$

其中$\rho = \phi_1 + \phi_2 + \cdots + \phi_{p-1}$。

$$\beta_j = -\phi_j + 1 - \cdots - \phi_p, j = 1,2,\cdots,p-1$$

如果$\phi_1 + \phi_2 + \cdots + \phi_p < 1$或$\rho < 1$，则序列平稳；

如果$\phi_1 + \phi_2 + \cdots + \phi_p = 1$或$\rho = 0$则序列非平稳。

则过程单位根的原假设为：$H_0 : \rho = 0$，ADF 检验统计量为：$\tau = \dfrac{\hat{p}}{SE\hat{p}}$。

3.2.2. ADF 检验的三种类型

(1) 无漂移项，无趋势项：$\nabla X_t = \rho X_{t-1} + \sum_{j=1}^{p-1} \beta_j \nabla X_{t-j} + w_t$；

(2) 有漂移项，无趋势项：$\nabla X_t = \alpha + \rho X_{t-1} + \sum_{j=1}^{p-1} \beta_j \nabla X_{t-j} + w_t$；

(3) 有漂移项，有趋势项：$\nabla X_t = \alpha + \beta_t + \rho X_{t-1} + \sum_{j=1}^{p-1} \beta_j \nabla X_{t-j} + w_t$。

检验模型的顺序为(3)、(2)、(1)。

3.3. 纯随机性检验

3.3.1. 纯随机序列定义

检验原理入下；

(1) 任意$t \in T$，有$EX_t = \mu$；

(2) 任意$t, s \in T$，有

$$\gamma(t,s) = \begin{cases} \sigma^2, t = s \\ 0, t \neq s \end{cases}$$

则称序列$\{X_t\}$为纯随机序列或白噪声序列。

3.3.2. 白噪声序列的性质

1) 纯随机性

$\gamma(k) = 0, \forall k \neq 0$，说明白噪声序列之间没有任何相关关系，如果$\gamma(k) \neq 0, \exists k = 0$则序列值之间呈现显著相关关系。

2) 方差齐性

时间序列分析中方差齐性是一个重要的性质，如果不满足方差齐性就说明该时间序列有异方差，所谓方差齐性即$DX_t = \gamma(0) = \sigma^2$。

3.3.3. 纯随机性检验原理

1) 假设条件

H0：$\rho_1 = \cdots = \rho_m = 0$，$\forall m > 1$；

H1：至少存在某个 $\rho_k \neq 0$，$\forall m > 1$，$k \leq m$。

2) 检验统计量

检验统计量如公式所示：

$$LB = n(n+2)\sum_{i=1}^{m}\left(\frac{\hat{\rho}_k^2}{n-k}\right) \sim x^2(m)$$

其中，n 为观测期数，m 为延迟期数。

3.4. ARIMA 模型参数估计的方法及原理

本文中初值是使用最小二乘法确定，ARIMA 的计算是用极大似然法。

3.4.1. 最小二乘估计

ARMA(p, q)模型，记：

$$\tilde{\beta} = \left(\phi_1 - \cdots - \phi_p, \theta_1 - \cdots - \theta_q\right)'$$

$$F_t\left(\tilde{\beta}\right) = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$$

残差项为：$\varepsilon_t = x_t - F_t\left(\tilde{\beta}\right)$

残差平方和为

$$Q\left(\tilde{\beta}\right) = \sum_{t=i}^{n}\left(x_t - \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}\right)^2$$

使残差平方和最小书估计 $\tilde{\beta}$ 前提。

3.4.2. 极大似然估计

$$L\left(\widehat{\beta_1}, \widehat{\beta_2}, \cdots, \widehat{\beta_k}; x_1, x_2, \cdots, x_n\right) = \max\left\{p\left(x_1, x_2, \cdots, x_n\right); \beta_1, \beta_2, \cdots, \beta_k\right\}$$

一般情况下时间序列所服从的正态分布是未知的，因而首先假定序列服从多元正态分布

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$$

其中 $\tilde{x} = \left(x_1, x_2, \cdots, x_n\right)$，$\tilde{\beta} = \left(\phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q\right)$，$\sum n = \left(\tilde{x}'\tilde{x}\right) = \Omega\sigma_\varepsilon^2$。

对数似然函数为：

$$\ln\left(\tilde{x}, \tilde{\beta}\right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\sigma_\varepsilon^2\right) - \frac{n}{2}|\Omega| - \frac{n}{2\sigma_\varepsilon^2}\ln\left[\tilde{x}'\Omega\tilde{x}\right]$$

求偏导，得到方程组：

$$\begin{cases} \dfrac{\partial}{\partial\sigma_\varepsilon^2}\ln\left(\tilde{x}, \tilde{\beta}\right) = -\dfrac{n}{2\sigma_\varepsilon^2} + \dfrac{s\left(\tilde{\beta}\right)}{2\sigma_\varepsilon^4} = 0 \\ \dfrac{\partial}{\partial\tilde{\beta}}\ln\left(\tilde{x}, \tilde{\beta}\right) = -\dfrac{1}{2}\dfrac{\partial\ln|\Omega|}{\partial\tilde{\beta}} + \dfrac{1}{\sigma_\varepsilon^2}\dfrac{\partial s\left(\tilde{\beta}\right)}{2\partial\tilde{\beta}} = 0 \end{cases}$$

式中，$s\left(\tilde{\beta}\right) = \tilde{x}'\Omega^{-1}\tilde{x}$。

3.5. 模型检验的方法与统计量

3.5.1. 模型的显著性检验

检验残差序列是否为白噪声。

H0： $\rho_1 = \cdots = \rho_m = 0$， $\forall m > 1$；

H1：至少存在某个 $\rho_k \neq 0$， $\forall m > 1$， $k \leq m$。

检验统计量如公式：

$$LB = n(n+2)\sum_{k=1}^{m}\left(\frac{\hat{\rho}_k^2}{n-k}\right) \sim x^2(m)。$$

如果检验后残差序列并不是白噪声，也就是说还有众多有效信息未被充分提取。

3.5.2. 模型参数的显著性检验

显著性检验的主要目的是检验模型中的参数是否显著非零，假设条件：

H0： $\beta_j = 0$；H1： $\beta_j \neq 0$， $\forall 1 \leq j \leq m$。

检验统计量： $T = \sqrt{n-m}\dfrac{\hat{\beta}_j - \beta_j}{\sqrt{a_{jj}Q(\hat{\beta})}} \sim t(n-m)$。

4. ARIMA 模型的建立

4.1. 数据来源

本文的数据的来源是国家统计局数据库年度数据以及网站 https://d.qianzhan.com/xdata/details/b25ae6e1c6e3c5c3.html。根据这两份数据算出 1978~2020 年的陕西全社会固定资产投资总额，单位为亿元。

4.2. 原时间序列的平稳性检验

根据图 1 所示，陕西全社会固定资产投资总额序列有明显的周期趋势、不在某个值附近随机波动、因此可以初步判定陕西全社会固定资产投资总额的原时间序列属于非平稳序列，需要对其进行数据的处理。

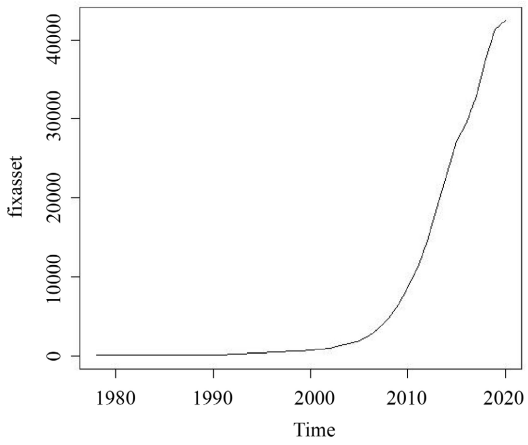

Figure 1. Trend chart of total fixed assets investment in Shaanxi from 1978 to 2020

图 1. 1978~2020 年陕西全社会固定资产投资总额趋势图

4.3. 数据处理

4.3.1. 平稳性检验

一阶差分后的时序图：从图 2 中我们可以看到陕西全社会固定资产投资总额还有明显的上升的。因此，需要二阶差分。二阶差分后的时序图图 3 中我们可以看到没有明显的上升的趋势。我们认为该平稳。平稳性检验结果显示，ADF 检验统计量的 P 值小于 0.05，所以可以确认二阶差分是一个平稳序列。

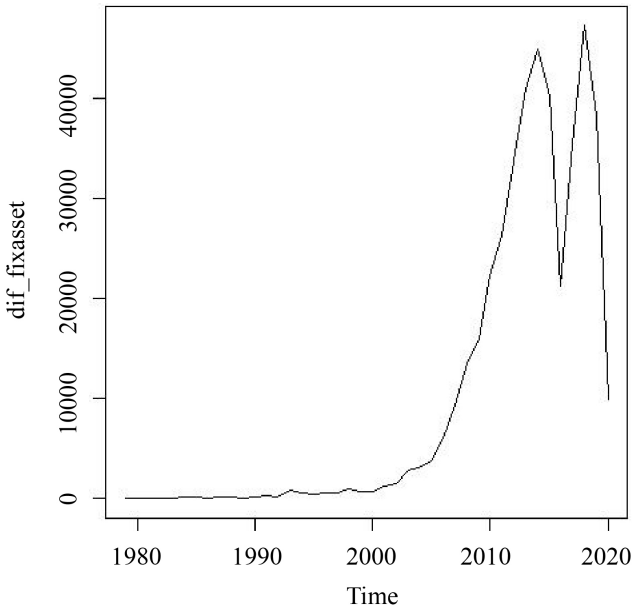

Figure 2. Sequence diagram after first order difference
图 2. 一阶差分后的时序图

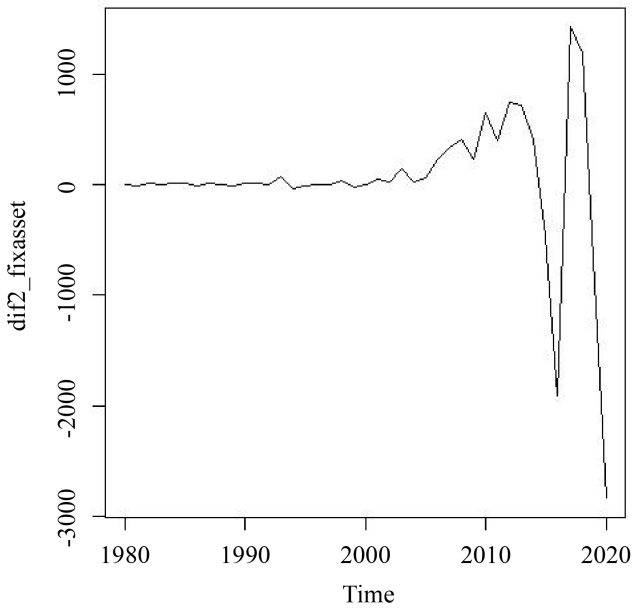

Figure 3. Sequence diagram after second order difference
图 3. 二阶差分后的时序图

4.3.2. 纯随机性检验

纯随机性检验是判断是否为平稳序列后一项重要步骤。结果如表 1 所示。

Table 1. Test results of pure random results
表 1. 纯随机结果检验结果

指标名称	延迟 6 阶	延迟 12 阶
X-SQUARED	16.69	18.806
P-VALUE	0.01049	0.09331

由表 1 纯随机检验结果显示，延迟 6 阶的统计量的 P 值小于 0.05，因此该序列为非白噪声序列。

4.4. 模型识别与定阶

差分后序列自相关图和偏自相关图如下图 4 和图 5 所示：

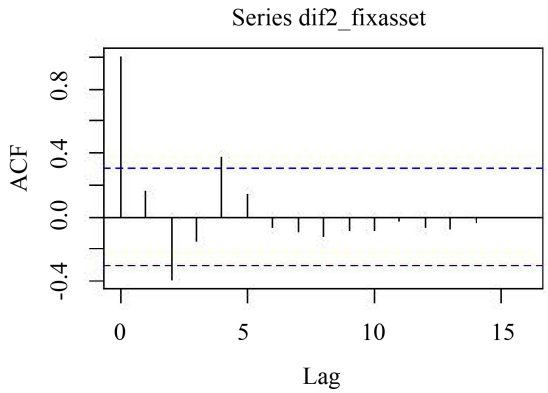

Figure 4. Sequence autocorrelation diagram after difference
图 4. 差分后序列自相关图

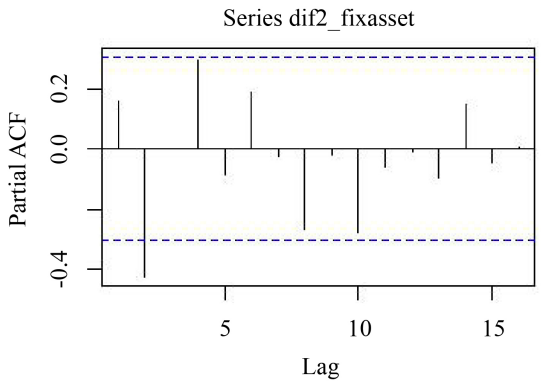

Figure 5. Sequence partial autocorrelation diagram after difference
图 5. 差分后序列偏自相关图

根据偏自相关图 5 所示，可以看出存在二阶截尾，而图 4 自相关图是五阶截尾，因此对 ARIMA(2,2,5)、

ARIMA(0,2,5)、ARIMA(2,2,0)进行拟合模型的比较，该模型均通过白噪声检验，根据 AIC 信息准则，我们得到最优模型，如图 6 所示：

```
Series: fixasset
ARIMA(0,2,1)

Coefficients:
          ma1
      0.5123
s.e.  0.1416

sigma^2 estimated as 393402:  log likelihood=-321.92
AIC=647.83   AICc=648.15   BIC=651.26
```

Figure 6. Optimal model
图 6. 最优模型

由图 6 可以看出，根据 AIC 信息准则我们得到的最优模型是 ARIMA(0,2,1)，则预测方程的表达式如下所示：$x_t = \varepsilon_t - 0.5123\varepsilon_{t-1}$。虽然根据 AIC 信息准则得到的最优模型是 ARIMA(0,2,1)，但是根据预测得到的信息可知，预测的结果不能很好地反映原序列的趋势，所以我们放弃这个模型。如下图 7 所示，根据 AIC 和 BIC 准则，我们最终选择模型 ARIMA(2,2,0)。

模型	AIC	BIC
ARIMA(2,2,5)	631.1455	644.8541
ARIMA(0,2,5)	633.4346	643.716
ARIMA(2,2,0)	640.279	645.4197

Figure 7. Comparison between AIC and BIC
图 7. AIC 和 BIC 比较

4.5. 模型的显著性检验

如图 8 所示，模型的显著性检验结果显示，残差序列可以视为白噪声序列。因此该模型显著成立。

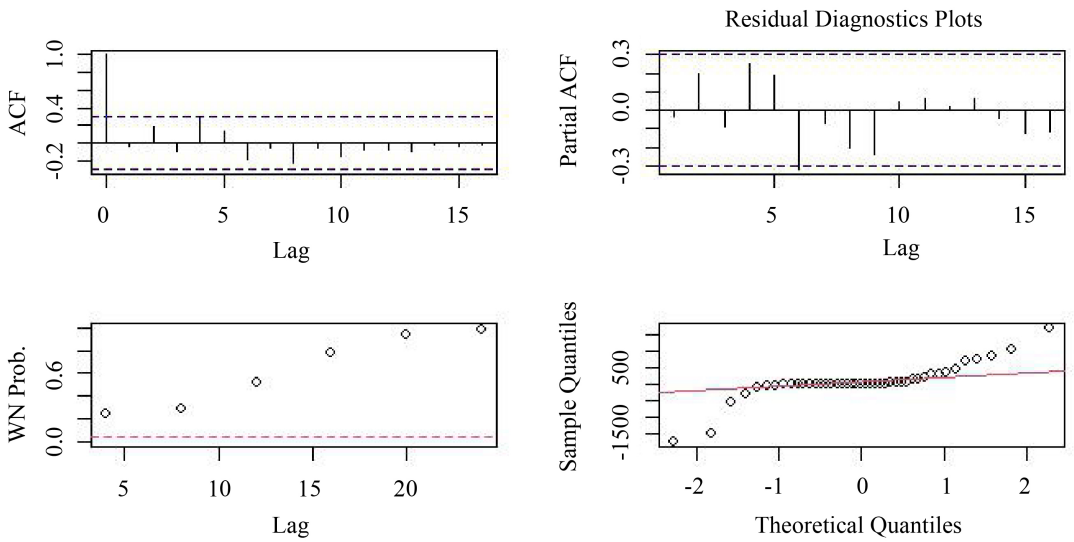

Figure 8. Model significance test
图 8. 模型的显著性检验

4.6. 模型的预测

根据拟合的模型 ARIMA(2,2,0)可预测出 2020~2030 年的的陕西全社会固定资产投资预测如下表 2 所示：

Table 2. Fixed assets investment model forecast of Shaanxi from 2021~2030
表 2. 2021~2030 年陕西全社会固定资产投资模型预测

年份	预测值
2021	43103.46
2022	45621.15
2023	48925.06
2024	51246.47
2025	52722.21
2026	54592.16
2027	57160.01
2028	59684.28
2029	61722.63
2030	63635.18

如图 9 所示可以看出实际数据和预测数据拟合的非常好。我们预测了后 10 年的数据。

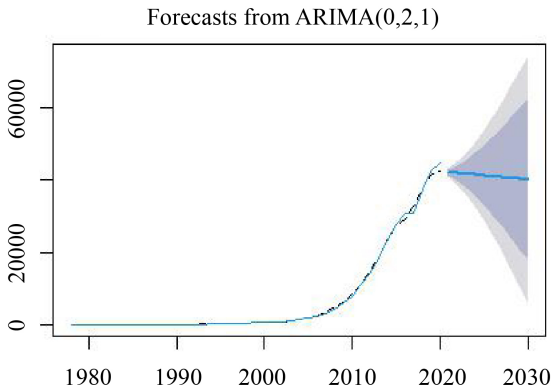

Figure 9. 2021~2030 Shaanxi social fixed asset investment forecast
图 9. 2021~2030 陕西全社会固定资产投资预测图

5. 结论与建议

本文对陕西全社会固定资产投资总额预测问题建立了 ARIMA 模型，年度陕西全社会固定资产投资总额数据呈现非平稳性，通过 2 阶差分后，陕西全社会固定资产投资总额呈现出平稳性，纯随机性检验后该序列为非白噪声序列，具有进一步研究的价值；模型的识别与定价过程中，通过不同模型的比较最终选择最优的 ARIMA(2,2,0)模型，模型的显著性检验结果也是显著的；2020 年之前的实际数据和预测数据拟合的非常好，显示出该模型的预测精度很高，具有很好的参考价值，因此进一步预测后十年的陕西

全社会固定资产投资总额。对陕西全社会固定资产投资总额的精准预测可以为根据预测结果制定相应的措施提供依据。所以，建议政府调整经济发展模式，完善监管制度，结合陕西的社会现实情况，在社会健康发展的基础上进行社会经济建设，促进社会经济健康稳定的发展。

参考文献

[1] 李惠. ARIMA 模型在我国全社会固定资产投资预测中的应用[J]. 黑龙江对外经贸, 2010(7): 87-88, 150.

[2] 靳宝琳, 赫英迪. ARIMA 模型在太原市全社会固定资产投资预测中的应用[J]. 太原科技大学学报, 2007, 28(5): 385-386.

[3] 石美娟. ARIMA 模型在上海市全社会固定资产投资预测中的应用[J]. 统计教育, 2004(3): 30-33.

[4] 舒服华, 马厚臣. 基于 Logistic 模型的河北省社会固定资产投资预测[J]. 保定学院学报, 2017, 30(6): 29-33, 40.

[5] 蒋艳, 庞林旗. 基于 Logistic 回归模型的北京市卫生总费用预测与发展阶段研究[J]. 中国社会医学杂志, 2021, 38(5): 580-582.

[6] 李俊松, 梁媛, 龙利兰. 成都平原经济区的物流需求预测研究[J]. 经贸实践, 2017(11): 1-4.

[7] 邓奎, 李龙国. 灰色新陈代谢模型在城市生活用水量预测中的应用[J]. 黑龙江水利科技, 2011, 39(1): 1-2.

[8] 马占青, 崔广柏, 杨宏杰, 王晓强. 城市污水排放的灰色马尔柯夫预测模型[J]. 河海大学学报(自然科学版), 2000(8): 1-3.

[9] 程欢, 姚建, 明星, 王沛. 等维动态递补灰色模型改进及应用研究[J]. 灌溉排水学报, 2016(5): 108-112.

[10] 丘健明, 王纯, 李斌. 全社会固定资产投资分析预测的多因素相关模型及优选研究[J]. 深圳大学学报(人文社会科学版), 2006, 23(6): 33-36.