Advances in Applied Mathematics 应用数学进展, 2023, 12(4), 1874-1883
Published Online April 2023 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2023.124194

Hans汉斯

基于ARIMA模型的深股指数预测分析

李　贺

云南财经大学统计与数学学院，云南　昆明

收稿日期：2023年3月26日；录用日期：2023年4月21日；发布日期：2023年4月28日

摘　要

股票是人们进行投资和理财的重要方式。随着我国人民生活水平的改善和提高，股票投资越来越受到人们的关注。对于股票的预测不仅能给投资者带来可观的收益，并且能很大程度上促进国民经济的发展。股票预测主要以预测股票未来的价格走势，帮助人们做出符合最大收益的决策为动力，具有较强的现实意义。本文选取了深股指数2019年12月30日到2022年11月18日的每日收盘指数，对数据进行平稳性检验后判断为非平稳序列。对数据进行差分化处理，再结合自相关性和偏自相关性的知识，结合差分、ACI和模型显著性的相关原则，构建了ARIMA模型，对深股指数进行总结和整体预测，为未来的相关研究提供一定的依据和参考。

关键词

股票，ARIMA模型，时间序列

Shenzhen Stock Index Predictive Analysis Based on ARIMA Model

He Li

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

Received: Mar. 26th, 2023; accepted: Apr. 21st, 2023; published: Apr. 28th, 2023

Abstract

Stocks are an important way for people to invest and manage their finances. With the improvement and improvement of the living standards of our people, stock investment has attracted more and more attention. The forecast of stocks can not only bring considerable returns to investors, but also greatly promote the development of the national economy. Stock forecasting is mainly driven by predicting the future price trend of stocks and helping people make decisions that are in line with the greatest returns, which has strong practical significance. In this paper, the daily clos-

文章引用: 李贺. 基于 ARIMA 模型的深股指数预测分析[J]. 应用数学进展, 2023, 12(4): 1874-1883.
DOI: 10.12677/aam.2023.124194

ing index of the Shenzhen stock index from December 30, 2019 to November 18, 2022 is selected, and the data is judged to be a non-stationary sequence after a stationary test. The data are differentiated, combined with the knowledge of autocorrelation and partial autocorrelation, combined with the correlation principles of difference, ACI and model significance, the ARIMA model is constructed, which summarizes and predicts the Shenzhen stock index, and provides a certain basis and reference for future related research.

Keywords

Stock, ARIMA Model, Time Series

1. 第 1 章绪论

1.1. 研究背景与意义

股票是公司在筹集资金时候的提供给出资人的一种证券，是投资者持有公司股份的证明。中国股市自 1989 年开始运行，经过三十多年的发展，现在已经有了上海、北京、深圳、香港、台湾五家证券交易所，也形成了较为完善的体系。股票如今在我国的地位越来越高，已经成为了人们进行投资理财的主要方式之一。

对于金融市场来说，收益和风险是成正相关关系的。股票高收益的同时也意味着较高的风险。在中国股票的市场上，个体散户仍然是主力军。由于散户很多不具备专业的知识储备和相关知识背景，使得他们的收益非常不稳定。因此，如何规避市场的高风险，同时为投资者带来收益，受到学者的广泛关注。

现如今，对于股票的研究已经有了一定的进展，但是由于股票的价格收到很多因素的影响，如公司的状况、投资者的情绪、经济发展状况、公共事件等都会影响到股票的未来走向，因此对于股票的预测研究仍然是一个具有挑战性的工作。

1.2. 文献综述

目前针对股票价格预测的相关研究已经有了大量的文献，不同学者利用不同的方法都对股票进行了相关的研究。主要大致可以分为：

利用传统逻辑回归预测方法进行分析预测，如逻辑回归、线性回归、GARCH 等等。Huang (2020)运用 Logistic 回归，通过筛选指标构建了股票预测模型，利用文本挖掘的技术将股票的非结构化数据量化，后基于情绪得分提出了 Logistic 回归的改进方法，为股票价格的预测精度提供了新的方向[1]。张少萍(2018)通过 Garch 模型对招商银行每日收盘价格进行技术分析，对其进行了预测，最后得到了较为理想的预测效果[2]。Zevallos (2019)通过分位数回归方法预测秘鲁股市的风险，得到了较为良好的效果[3]。

股票作为时间序列的数据，国内外也有很多学者利用 ARIMA 模型对股票价格进行预测。Chi (2018)通过对短期股票的分析预测，运用 ARIMA(1,2,2)模型，为投资者提供了可靠的信息，在一定程度上为投资者的决策提供了相关的参考[4]。黄莉霞(2020)通过对中国平安的分析，有别于以往，从市盈率出发，构建了 ARIMA(0,1,1)模型，对中国平安的股票进行 5 日内的市盈率预测，结果显示误差只在 5%左右，使股票的市盈率有了很好的预测模范[5]。赵庆国(2020)则说明时序权重均值模型是通过利用时间序列的

相关知识，对不同时期的价格赋予不同的权重，以此减轻了滞后性，更好地反映出了价格走向[6]。张颖超(2019)通过对上证指数的分析预测，准确的计算了预测误差，最后选定了 ARIMA(4,1,4)模型，证明了该模型可以预测短期数据[7]。

除此之外，很多学者在股票价格预测模型中引入了灰色关联、神经网络等其他方法，以此来提高模型预测的精度和准度。王振兴(2011)通过建立 BP-RBF 双神经网络模型，将组合模型和单一模型进行对比，对我国的股市进行分析预测，提高了预测的准确率[8]。杨琦等(2016)构建了基于 ARMA-GARCH 的组合模型，用 GARCH 去除了 ARMA 模型产生的异方差，达到了较好的短期价格预测能力，并且验证了 ARMA-GARCH 模型的可行性[9]。姜乐(2015)分别用传统的时间序列模型和 BP 神经网络模型对股票进行了价格预测，发现这两种方法可以预测股票价格的变化趋势，然后构建了 ARMA-BP 模型提高了预测的精度[10]。

综合以上文献，目前学者对于股票价格预测有多种方法，但是选取的数据往往是移动均线数据，而非每天的收盘价。由于移动均线数据具有一定的滞后效应，因此会对模型预测的精度产生影响。且部分学者预测对象是个股而非整体股市，由于股票之间存在的差异性较大，因此预测个体股票的指导效果并不强。本文选取了深股指数 2019 年 12 月 30 日至 2022 年 11 月 18 日每日的收盘指数为样本，利用 R 语言构建 ARIMA 模型对对股市未来走向进行了预测，并给出指导意见。

2. 第 2 章理论基础

2.1. 平稳性时间序列

2.1.1. 平稳时间序列

平稳性时间序列根据限制性条件的严格程度的不同，分为严平稳时间序列和宽平稳时间序列两种定义。严平稳是限制条件比较苛刻的一种，将序列所有的统计性质都不随着时间的推移而变化的序列定义为平稳序列；宽平稳则认为序列的统计性质主要由它的低阶矩决定，只要保证序列的低阶矩平稳，就能保证序列的主要性质近似平稳。在实际应用中，如果不特别注明为严平稳，那么就泛指宽平稳，本文所使用的序列平稳也为宽平稳。宽平稳时间序列$\{X_t\}$满足以下三个条件：

任取$t \in T$，有$EX_t^2 < \infty$；

任取$t \in T$，有$EX_t = \mu$，μ为常数；

任取$t,s,k \in T$，且$k+s-t \in T$，有$\gamma(t,s) = \gamma(k,k+s-t)$。

2.1.2. 平稳性检验

序列的平稳性检验有两种方法：一种是根据时序图和自相关图做出判断的图检验法，另一种是构造检验统计量进行假设检验的方法。图检验法具有一定的主观色彩，而统计检验方法能在一定的可靠性水平之下对序列的平稳性做出判别，使检验更加可信。

平稳性统计检验方法的理论思想：如果序列是平稳的，那么该序列的所以特征根都在单位圆内，反之为非平稳序列。因此，平稳性统计检验方法又叫作单位根检验。单位根检验有诸多方法，本文使用的是较为经典是 ADF 检验。ADF 检验的假设条件如下：

原假设 H0：序列存在单位根，为非平稳序列；

备择假设 H1：序列不存在单位根，为平稳序列；

ADF 检验统计量：

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

式子中，$S(\hat{\rho})$为$\hat{\rho}$的样本标准差。当τ统计量小于α分位点，或者τ统计量的 P 值小于显著性水平α

时，可认为序列平稳，反之不平稳。

2.2. 非平稳时间序列

ARIMA 模型又称为差分自回归移动平均模型。ARIMA(p,d,q)模型中，AR 是“自回归”，p 为自回归项数；MA 为“移动平均”，q 为移动平均项数，d 为使之成为平稳序列所做的差分阶数。

果一个时间序列$\{Y_t\}$的 d 次差分 $W_t=\nabla^d Y_t$ 是一个平稳 ARMA 的过程，我们称$\{Y_t\}$为自回归滑动求和平均模型，如果 W_t 服从 ARMA(p,q)模型，则称$\{Y_t\}$是 ARIMA(p,d,q)过程。

一般的表达方式为：

$$\begin{cases} W_t=\phi_1 W_{t-1}+\phi_2 W_{t-2}+\cdots+\phi_p W_{t-p}+\varepsilon_t-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-\cdots-\theta_q\varepsilon_{t-q} \\ \phi_p\neq 0,\theta_q\neq 0 \\ E\left(W_t\varepsilon_t\right)=0,\forall s<t \\ \varepsilon_t\sim WN\left(0,\sigma^2\right),\forall s<t \end{cases}$$

其中，$\phi_1,\cdots,\phi_p,\theta_1,\cdots,\theta_q$ 为模型系数，$\nabla Y_t=Y_t-Y_{t-1},\nabla^2 Y_t=\nabla Y_t-\nabla Y_{t-1},\cdots$，$W_t=\nabla^d Y_t=\nabla^{d-1}Y_t-\nabla^{d-1}Y_{t-1}$，当 $d=1$ 时，$W_t=Y_t-Y_{t-1}$，则有

$$W_t=\phi_1 W_{t-1}+\phi_2 W_{t-2}+\cdots+\phi_p W_{t-p}+\varepsilon_t-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-\cdots-\theta_q\varepsilon_{t-q}$$

用观测序列表示为：

$$\begin{aligned} Y_t-Y_{t-1}&=\phi_1\left(Y_{t-1}-Y_{t-2}\right)+\phi_2\left(Y_{t-2}-Y_{t-3}\right)+\cdots+\phi_p\left(Y_{t-p}-Y_{t-p-1}\right)+\varepsilon_t-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-\cdots-\theta_q\varepsilon_{t-q} \\ &=\phi_1 Y_{t-1}+\left(\phi_2-\phi_1\right)Y_{t-2}+\cdots+\left(\phi_p-\phi_{p-1}\right)Y_{t-p}-\phi_p Y_{t-p-1}+\varepsilon_t-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-\cdots-\theta_q\varepsilon_{t-q} \\ Y_t&=\left(\phi_1+1\right)Y_{t-1}+\left(\phi_2-\phi_1\right)Y_{t-2}+\cdots+\left(\phi_p-\phi_{p-1}\right)Y_{t-p}-\phi_p Y_{t-p-1}+\varepsilon_t-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-\cdots-\theta_q\varepsilon_{t-q} \end{aligned}$$

上式称为模型的差分形式，我们假设模型自 $t=-m$ 开始，得到 ARIMA(p,1,q)表达式为：

$$Y_t=\sum_{j=-m}^{t}W_j$$

通过两次求和处理 ARIMA(p,2,q)可以得到：

$$Y_t=\sum_{j=-m}^{t}\sum_{i=-m}^{j}W_i=\sum_{j=0}^{1+m}(j+1)W_{t-j}$$

实际中，d 通常取 1 或者 2。

2.3. 纯随机性检验

2.3.1. 纯随机序列

纯随机序列，又叫白噪声序列，白噪声序列的各项序列值之间没有任何相关关系，在进行完全无序的随机波动，具有以下性质：

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

一旦某个序列表现出纯随机性质，那么该随机事件就不包括任何有用信息值得提取，此时对该时间序列的分析就该终止。

2.3.2. 纯随机检验

纯随机性检验也称白噪声检验，是检验序列是否为纯随机序列的一种方法。由于序列之间的变异性

是绝对的，相关性是偶然的，因此，序列纯随机性检验的假设条件如下：

原假设 H0：延迟期数小于或者等于 m 期的序列之间相互独立；

备择假设 H1：延迟期数小于或者等于 m 期的序列之间有相关性。

LB(Ljung-Box)统计量：

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

式中，n 为序列观察期数，m 为指定延迟期数，且 LB 统计量近似服从自由度为 m 的卡方分布。

2.4. 模型显著性检验

在拟合完模型之后需要对模型的显著性进行检验，检验模型是否有效，是否充分提取了时间序列的信息。对模型的检验实际上是对模型的残差序列来进行相关性的检验，一个好的拟合模型应该能够提取观察序列中几乎所有的样本相关信息。因此，模型显著性检验就可以转化为模型拟合残差项的白噪声检验，即对拟合残差构建 LB 统计量(理论和前文一致)，进行检验。拟合残差序列为白噪声序列，就称模型显著性有效，若拟合残差序列为非白噪声序列，则拟合模型不够有效，需选择其他模型拟合。

2.5. 参数显著性检验

模型参数的显著性检验就是要检验模型中的每一个未知参数是否显著，是模型更精简。如果模型中的某个参数不显著非零，则该模型就应该剔除该自变量。模型参数的显著性检验的假设条件如下：

原假设 H_0：$\beta_j=0$；

备择假设 H_1：$\beta_j\neq 0,\forall 1\le j\le m$；

检验统计量 t 为：

$$T=\sqrt{n-m}\,\frac{\hat{\beta}_j}{\sqrt{\alpha_{jj}Q\left(\tilde{\beta}\right)}}$$

上式中，T 服从自由度为 n-m 的 t 分布。当该检验统计量的 P 值小于显著性水平 α 时，拒绝原假设，认为该参数显著非零。反之，不能拒绝原假设，应该剔除不显著参数。

3. 第 3 章 ARIMA 模型的建立与预测

3.1. 数据来源

本文所选取的数据是 2019 年 12 月 30 日~2022 年 11 月 18 日的深证日指数，共 700 条数据，所有的数据均可以从官网得到。

3.2. 数据预处理

首先将数据录入到软件 R 中，画出该数据的时序图，对数据的平稳性进行初步的判断，结果如图 1 所示。从图中可以看出该序列具有明显的上升和下降趋势，初步判断该序列为非平稳的。但是由于图示带有主观倾向，因此对其进行 ADF 检验之后，显示为显著，则拒绝原假设，认为序列为非平稳的。

对该序列进行一阶差分，差分后的序列时序图见图 2。显示差分后的序列已不在具有明显的趋势，因此认定该序列具有平稳的特征 ADF 结果显示(见表 1)无截距无趋势、有截距无趋势、有截距有趋势三种情况的 ADF 的 P 值均小于显著性水平 0.05，证明该序列经过差分后变的平稳。对该序列进行纯随机性检验，结果如表 2 所示，P 值小于显著性水平 α，因此拒绝原假设，该序列非白噪声，可以进行分析。

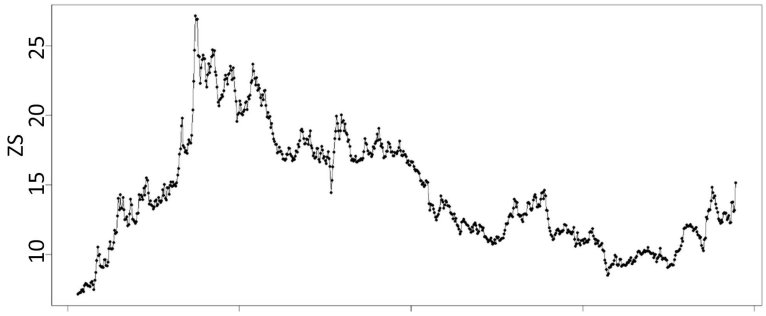

Figure 1. Shenzhen stock index time series chart
图 1. 深股指数时序图

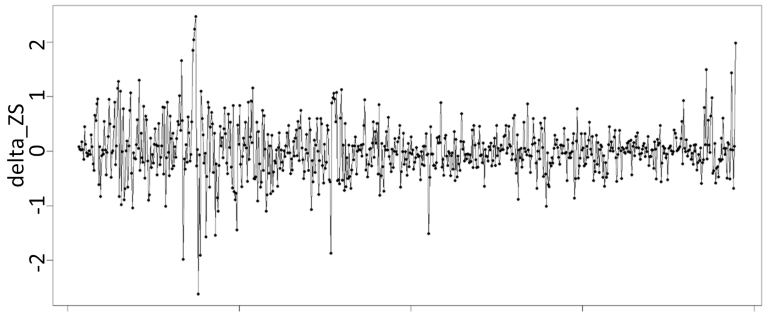

Figure 2. Differential timing chart of the Shenzhen stock index
图 2. 深股指数差分后时序图

Table 1. The ADF value after the first-order differential of the Shenzhen stock index
表 1. 深股指数一阶差分后 ADF 值

无截距无趋势			有截距无趋势			有截距有趋势		
Lag	ADF	P.value	Lag	ADF	P.value	Lag	ADF	P.value
0	−23.1	0.01	0	−23.1	0.01	0	−23.1	0.01
1	−16.2	0.01	1	−16.2	0.01	1	−16.3	0.01
2	−15.1	0.01	2	−15.1	0.01	2	−15.1	0.01
3	−13.4	0.01	3	−13.4	0.01	3	−13.4	0.01
4	−12.8	0.01	4	−12.8	0.01	4	−12.8	0.01
5	−11.6	0.01	5	−11.6	0.01	5	−11.7	0.01
6	−10.7	0.01	6	−10.7	0.01	6	−10.7	0.01

Table 2. Results of pure random test of the series after the difference of the Shenzhen stock index
表 2. 深股指数差分后序列纯随机检验结果

滞后阶数	X-squared	p-value
1	9.5042	0.00205

3.3. 拟合模型

对序列进行差分之后的自相关性和偏自相关性如图 3 所示。自相关图一阶拖尾，偏自相关图 3 阶拖尾，可以拟合模型 ARIMA(1,1,1)、ARIMA(0,1,1)、ARIMA(1,1,0)、ARIMA((1,3),1,0)。

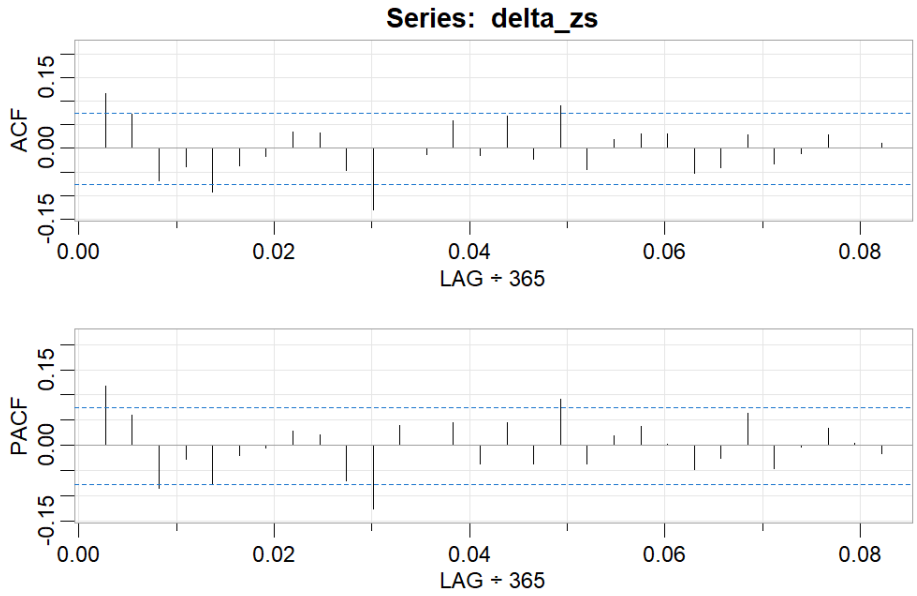

Figure 3. Autocorrelation plots and partial autocorrelation plots of the first-order difference of the Shenzhen stock index
图 3. 深股指数的一阶差分的自相关图和偏自相关图

根据对以上自相关图和偏自相关图的特征，尝试对以下模型进行拟合，如表 3 所示。

Table 3. Model ordering and model selection
表 3. 模型的定阶及模型选择

模型	自相关函数	偏自相关函数	拟合模型
模型 1	拖尾	拖尾	ARIMA(1,1,1)
模型 2	1 阶截尾	拖尾	ARIMA(0,1,1)
模型 3	拖尾	1 阶截尾	ARIMA(1,1,0)
模型 4	拖尾	1、3 阶截尾	ARIMA((1,3),1,0)

根据表 3，用 R 拟合上述模型，得到拟合结果如表 4 所示。

Table 4. The results of the model fit
表 4. 各模型拟合结果

模型	AIC
模型 1	997.18
模型 2	997.46
模型 3	996.16
模型 4	993.69

由表 3 中模型的拟合结果，可以看到模型 4 的 AIC 值最小，因此在拟合的四个模型中，ARIMA((1,3),1,0)模型是最优的。为了避免模型定阶的主观性，再采用系统自动定阶的语句建立 ARIMA 模型，得到的模型为 ARIMA(1,1,3)。ARIMA 模型的 AIC 为 991.84，为所有模型中最低，因此暂定为 ARIMA(1,1,3)对模型进行显著性检验，结果如图 4 所示。

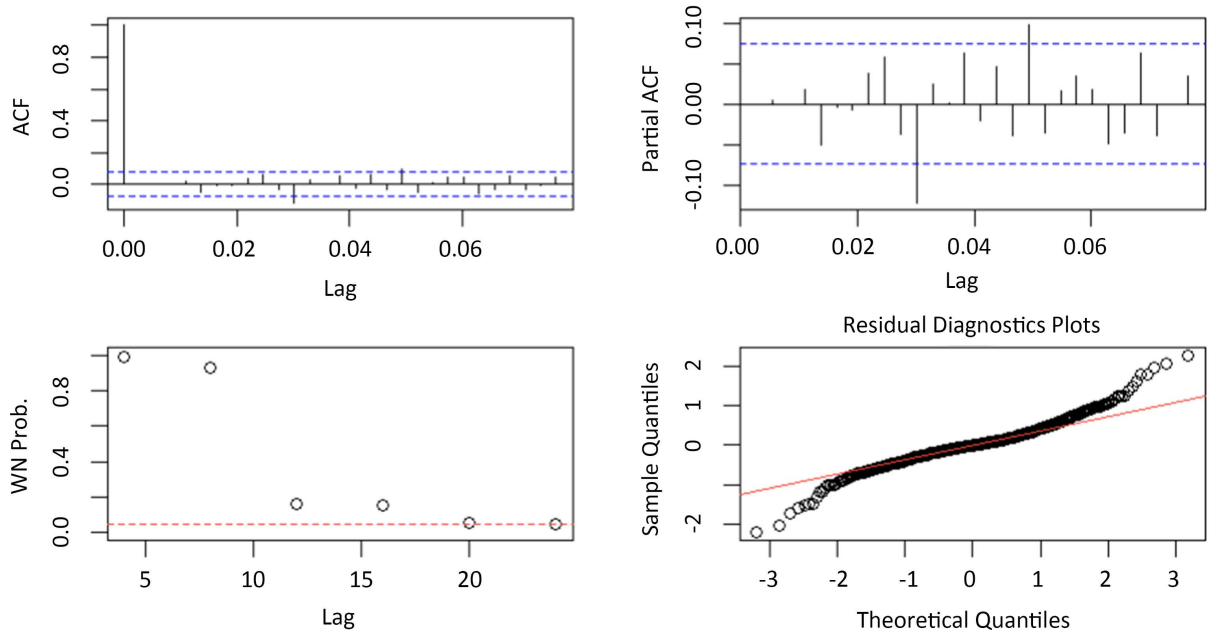

Figure 4. Test plot for significance of the DIF sequence fitting model
图 4. dif 序列拟合模型显著性检验图

对模型的参数显著性进行检验，结果如表 5。

Table 5. Automatic ordering model results
表 5. 自动定阶模型结果

Ar1	Ma1	Ma2	Ma3	AIC
0.7431	-0.6324	-0.759	-0.1233	991.84
0.1303	0.1336	0.0457	0.0365	

模型为：$\widehat{y_t} = \mu + 0.7431y_{t-1} - 0.6324e_{t-1} - 0.759e_{t-2} - 0.1233e_{t-3}$。

根据结果判断，参数的估计值大于它的两倍标准差则认为该模型可以有效。且 ARIMA(1,1,3)的 AIC 效果最好，所以综合选择 ARIMA(1,1,3)模型作为预测模型。

3.4. ARCH 检验

通过对模型 ARIMA(1,1,3)进行 ARCH 检验，结果如表 6 所示。

Table 6. ARCH test results
表 6. ARCH 检验结果

X-squared	df	P-value
3.3078e-06	1	0.9985
0.013982	2	0.993
0.014162	3	0.9996
0.25066	4	0.9928
1.9998	5	0.8492
2.011	6	0.9187

由图已知，由于 p 值均大于 0.05，因此接受原假设，认为该模型不存在异方差。

3.5. 模型预测

对 ARIMA(1,1,3)模型对后 10 天的深股指数进行预测，我们得到了结果如图 5 所示。

```
Forecast for univariate time series:
    Lead Forecast   S.E Lower Upper
701    1     15.4 0.488  14.5  16.4
702    2     15.6 0.730  14.1  17.0
703    3     15.4 0.931  13.6  17.2
704    4     15.3 1.078  13.2  17.4
705    5     15.2 1.197  12.9  17.6
706    6     15.2 1.297  12.6  17.7
707    7     15.1 1.385  12.4  17.8
708    8     15.1 1.465  12.2  17.9
709    9     15.1 1.538  12.0  18.1
710   10     15.0 1.606  11.9  18.2
```

Figure 5. ARIMA (1,1,3) model prediction results
图 5. ARIMA(1,1,3)模型预测结果

对预测数据可视化，得到图 6。

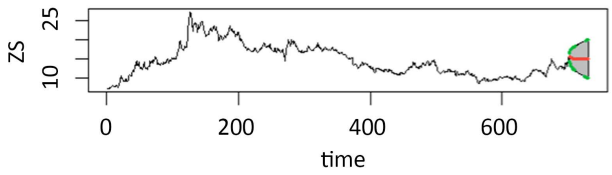

Figure 6. Shenzhen stock index forecast chart
图 6. 深股指数预测图

Table 7. Comparison of prediction results
表 7. 预测结果对比

时间	预测值	实际值	相对误差
2022/11/21	15.4	15.22	1.18%
2022/11/22	15.6	15.41	1.23%
2022/11/23	15.4	15.31	0.5%
2022/11/24	15.3		
2022/11/25	15.2		
2022/11/28	15.2		
2022/11/29	15.1		
2022/11/30	15.1		
2022/12/1	15.1		
2022/12/2	15.0		

由表 7 可知，该序列预测模型与原图一样具有稳定趋势，因此预测结果较好。

4. 结论

本文针对 2019 年 12 月到 2022 年 11 月的深股指数进行了预测分析。历史收盘指数为非平稳序列，通过一阶差分使其变为平稳序列。经过纯随机检验认为差分平稳序列非白噪声，有进一步建立模型的意义。接着通过观察一阶差分的自相关图和偏自相关图拟合了不同的模型，接着通过 AIC 最小原则确定了模型为 ARIMA(1,1,3)，拟合的模型通过了模型显著性检验和参数显著性检验。通过 ARIMA(1,1,3)模型对未来十天的股市走向进行了预测和分析，结果符合以往的发展趋势，具有较强的参考意义。

参考文献

[1] Huang, J.-Y. and Liu, J.-H. (2020) Using Social Media Mining Technology to Improve Stock Price Forecast Accuracy. *Journal of Forecasting*, **39**, 104-116.3. https://doi.org/10.1002/for.2616

[2] 张少萍. 招商银行股票价格预测研究——基于 GARCH 模型[J]. 金融经济, 2018(10): 84-87.

[3] Zevallos, M. (2019) A Note on Forecasting Daily Peruvian Stock Market Volatility Risk Using Intraday Returns. *Economia*, **42**, 94-101. https://doi.org/10.18800/economia.201902.004

[4] Chi, W.L. (2018) Stock Price Forecasting Based on Time Sseries Aanalysis. *AIP Conference Proceedings*, **1967**, 040032. https://doi.org/10.1063/1.5039106

[5] 黄莉霞. 基于 ARIMA 模型的股价分析与预测——以中国平安为例[J]. 科技经济市场, 2020(10): 62-63.

[6] 赵庆国, 孔祥月, 刘莉明, 杨龙倩. 短期股票价格预测的时序权重均值模型构建[J]. 沈阳航空航天大学学报, 2020, 37(4): 81-89.

[7] 张颖超, 孙英隽. 基于 ARIMA 模型的上证指数分析与预测的实证研究[J]. 经济研究导刊, 2019(11): 131-135.

[8] 王振兴. BP-RBF 组合神经网络在股票预测中的应用研究[D]: [硕士学位论文]. 兰州: 兰州商学院, 2011.

[9] 杨琦, 曹显兵, 基于 ARMA-GARCH 模型的股票价格分析与预测[J]. 数学的实践与认识, 2016, 46(6): 80-86.

[10] 姜乐. 基于时间序列的股票价格分析研究与应用[D]: [硕士学位论文]. 大连: 大连理工大学, 2015.