ARIMA 模型

第 31 组　罗爽 张宗璞 张家儒

一、模型概览

ARIMA 模型全称为差分自回归移动平均模型 (Autoregressive Integrated Moving Average Model, 简记 ARIMA 模型)，是由博克思 (Box) 和詹金斯 (Jenkins) 于 70 年代初提出的一著名时间序列预测方法，所以又称为 box-jenkins 模型、博克思 -詹金斯法。其中 ARIMA（p，d，q）称为差分自回归移动平均模型，AR 是自回归, p 为自回归项; MA 为移动平均，q 为移动平均项数，d 为时间序列成为平稳时所做的差分次数 [1]。

ARIMA 模型的基本思想是：将预测对象随时间推移而形成的数据序列视为一个随机序列，用一定的数学模型来近似描述这个序列，这个模型一旦被识别后就可以从时间序列的过去值及现在值来预测未来值。其主要应用在统计学和计量经济学中，特别是在时间序列分析中，之后又根据不同的应用场景发展出了一些变体，如多个时间序列的 VARIMA 模型，考虑了季节因素的 SARIMA 模型，针对时间序列表现出了长范围依赖性的 FARIMA 或 ARFIMA 模型 [2]。

二、模型介绍

ARIMA 模型中的参数取不同的值，可以构成不同的模型，如自回归模型 (Autoregressive model，简称 AR 模型)，移动平均模型 (Moving Average model, 简称 MA 模型)，自回归移动平均模型 (Autoregressive Moving Average model, 简称 ARMA 模型), 下面将分别介绍这几个模型。

2.1　AR 模型 [3]

AR 模型广泛应用在经济学、资讯学、自然现象的预测上，此模型是用同一变量的历史数据来预测现在或未来数据的模型，即用 x_1 至 x_{t-1} 来预测 x_t，并假设它们之间为线性关系，所以称为自回归，模型假设的线性关系如下式：

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t \tag{1}$$

其中 c 为常数，φ_i 为自相关系数，p 为 AR(p) 模型的阶数，ε_t 假设为平均数为 0，标准差为 σ 的随机误差值，σ 被假设为对任何的都不变。

2.2 MA 模型 [3]

AR 模型描述的是当前值与历史值之间的关系，而 MA 模型描述的是自回归部分的误差累计, 当前预测值是历史白噪声的线性组合，其数学表达如下：

$$X_t = \mu + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{2}$$

其中 μ 是时间序列的均值，q 是 MA(q) 模型的阶数，θ_i 是参数，ε_t 表示白噪声。

2.3 ARMA 模型 [3]

ARMA 模型是 AR 模型和 MA 模型的结合，常用于市场研究中，如消费行为模式的变迁，销售量、市场规模的预测等，ARMA(p,q) 模型包含了 p 个自回归项和 q 个移动平均项，可以表示为：

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{3}$$

2.4 ARIMA 模型 [2]

AR 模型、MA 模型和 ARMA 模型均适用于平稳的时间序列，当原始的时间序列为非平稳时间序列时，需要先进行一步差分操作，便产生了 ARIMA(p,d,q) 模型，即在 ARMA 模型的基础上多了差分操作，即参数 $d > 0$。ARIMA(p,d,q) 模型可以表示为：

$$(1 - \sum_{i=1}^{p} \varphi_i L^i)(1 - L)^d X_t = (1 + \sum_{i=1}^{q} \theta_i L^i)\varepsilon_t \tag{4}$$

其中 L 是滞后算子，$L^i X_t = X_{t-i}$。

2.5 使用模型的相关背景知识

时间序列的平稳性 [4]

若某个时间序列由某一随机过程生成，其均值和方差都是与时间无关的常数，两个时间点数据的协方差是只与时间间隔有关，与时间无关，则称由该随机过程生成的时间序列是弱平稳的。在实际应用中，一个时间序列满足弱平稳的条件，则认为是平稳的，即在数据的时间图上显示观测数据在一个常数水平上下以相同幅度波动。

白噪声序列

白噪声序列是白噪声过程的样本实称，白噪声过程是指期望为 0，方差为常数且随机变量之间互不相关的随机过程。对白噪声序列，所有自相关函数为 0。

自相关与偏自相关 [4]

自相关函数是一个信号与其自身在不同时间点的互相关，表达了同一过程不同时刻的相互依赖关系，同一时间序列时间 s 和时间 t 的自相关定义为：

$$R(s,t) = \frac{E[(X_t - \mu_t)(X_s - \mu_s)]}{\sigma_t \sigma_s} \tag{5}$$

其中 E 表示期望，μ 表示均值，σ 表示方差，在弱平稳条件下，自相关系数只与时间间隔有关，而与具体的时间无关。

偏自相关是指同一时间序列中的 X_t 与 X_{t-k} 在排除了两者之间的变量 X_{t-1},X_{t-2},...,X_{t-k+1} 影响之后的互相关。

三、模型应用

3.1 使用方法 [1]

使用 ARIMA 模型对时间序列进行预测的过程如下：

(1) 根据时间序列的散点图、自相关函数和偏自相关函数图识别其平稳性。

(2) 对非平稳的时间序列数据进行平稳化处理。直到处理后的自相关函数和偏自相关函数的数值非显著非零。

(3) 根据所识别出来的特征建立相应的时间序列模型。平稳化处理后，若偏自相关函数是截尾的，而自相关函数是拖尾的，则建立 AR 模型；若偏自相关函数是拖尾的，而自相关函数是截尾的，则建立 MA 模型；若偏自相关函数和自相关函数均是拖尾的，则序列适合 ARMA 模型。使用自相关函数和偏自相关函数识别模型的方式总计为下表：

表 1　ARMA 模型相关图特征

模型	自相关系数	偏自相关系数
AR(p)	拖尾	p 阶截尾
MA(q)	q 阶截尾	拖尾
ARMA(p,q)	拖尾	拖尾

(4) 参数估计，检验是否具有统计意义。

(5) 假设检验，判断（诊断）残差序列是否为白噪声序列。

(6) 利用已通过检验的模型进行预测。

3.2 应用举例 –基于 ARIMA 模型的备件消耗预测 [5]

应用背景

在军事活动中，快速响应战场需求是装备战斗力的重要指标之一，而快速响应战场需求，部队需要保证有一定数量备件，实际中却常常由于没有足够的备件导致装备不能快速形成战斗力，造成备件短缺的重要原因就是使用的备件需求预测方法和模型不够精确，因此需要找到一个比较精确的预测模型，此处使用 ARIMA 模型对备件消耗进行预测。

应用过程

(1) 基本思路：将备件消耗的历史统计数据视为一个时间序列，即为一组依赖于时间 t 的随机变量序列，变量间有县官性，并表现出一定的规律性，根据这些历史消耗数据尽可能建立合理的统计模型，用模型解释规律性并进行预测。

(2) 原始数据及处理：以航空兵场站某种航材备件 3 年的消耗率（件/1000h) 进行分析和预测，取前 30 组数据建立模型，并用后面的几组数据对模型进行预测验证。2001 年 1 月到 2003 年 12 月三年的原始数据时间序列如图1：从图1可以看出数据有明显递

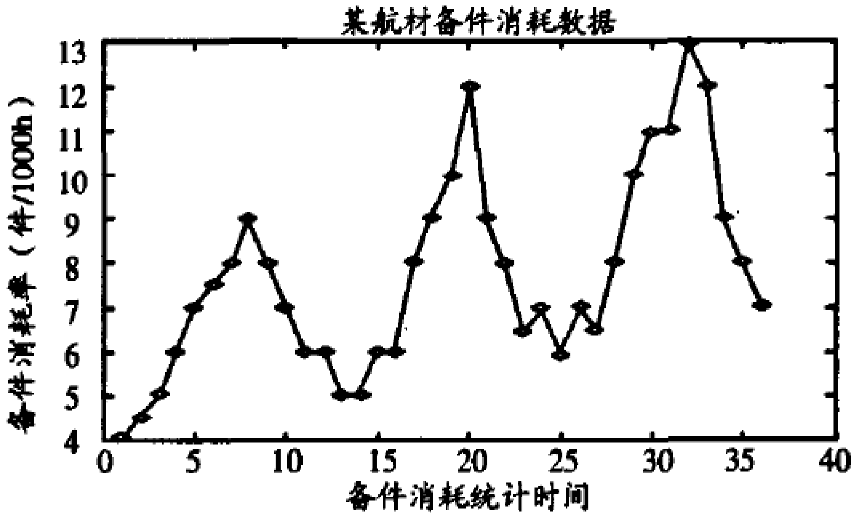

图 1　某航材备件消耗数据时间序列图

增的趋势，为非平稳序列，尝试一次差分做平稳化处理，结果仍未平稳，再进行 2 次差分，进行游程检验，可以通过检验，认为 2 次差分后的数据平稳，即 ARIMA(p,d,q) 中的参数 d 确定为 2，并将数据零均值化。

(3) 确定模型阶数：计算零均值化后的序列的自相关函数和偏自相关函数，结果如图2。其中上下两条线为置信区间，由图2可以看出 $0 \leq p \leq 3, 0 \leq q \leq 2$。对 p、q 可能的组合进行参数估计，并使用 AIC 准则进行定阶，对估计出的参数进行平稳性和可逆性检验，初步确定最佳模型为 ARIMA(3,2,1) 模型。

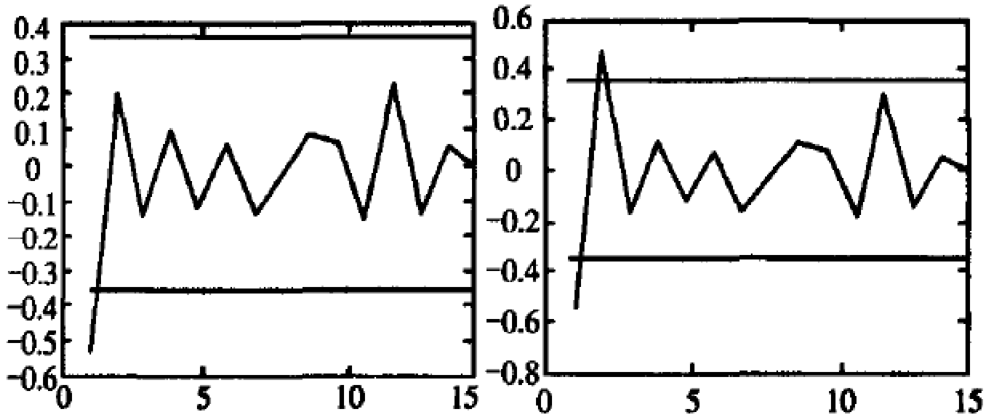

图 2　新序列的 ACF（左）和 PACF（右）

(4) 白噪声检验：对通过平稳性和可逆性检验的模型进行白噪声检验，通过检验，确认模型。

(5) 进行预测：使用一步预测法进行预测，预测结果如图3。

时间	真实值	移动平均法[5]				ARIMA 模型			
		预测值	MAE	MRE	MSE	预测值	MAE	MRE	MSE
2003.09	12	7.2447	2.0621	21.83%	1.3524	13.4777	0.6922	6.93%	0.4298
2003.10	9	8.7642				9.8088			
2003.11	8	8.8250				7.7322			
2003.12	7	9.4324				6.7857			

图 3　预测结果对比

3.3　应用举例 –基于 ARIMA 模型的上证指数走势预测 [6]

本实例来自于 2016 江西财经大学数模竞赛中的一份答卷。

应用背景

随着我国日渐成为 21 世纪最重要的国家，国内股票市场的波动，牵动着亿万投资者的心弦，先看两张上证综指的大盘走势图，如图4和图5, 图4是 2014 年 1 月 -2016 年 3 月的上证综指，图5是 2015 年 1 月 -2010 年 12 月的上证综指，将两图进行对比，发现 2016 年 3 月的市场与 2009 年的市场都经历了快速上涨之后的调整，估值也都处于历史中等偏低水平，注意到 2016 年 3 月与 2009 年相同的以稳增长为主的政策环境，同样曾经经历了大宗商品较大幅度的下跌、投资者关于人民币汇率贬值及经济前景偏于悲观的类似预期，有人认为 2016 年的市场状况类似于 2009 年。本例将用上证综指 2005

年 1 月 -2016 年 3 月的逐日数据建模，从市场走势角度考证上述“历史类似性”，并对未来一周及未来一月的上证综指走势进行预测。

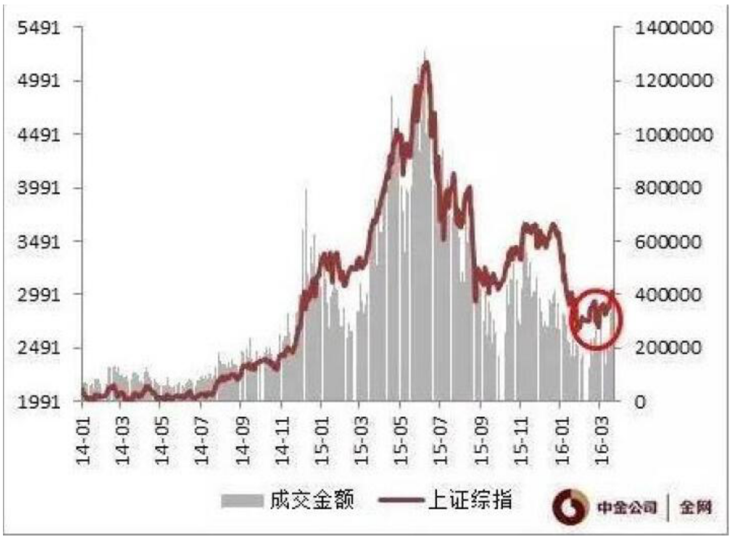

图 4　2014 年 1 月 -2016 年 3 月上证综指

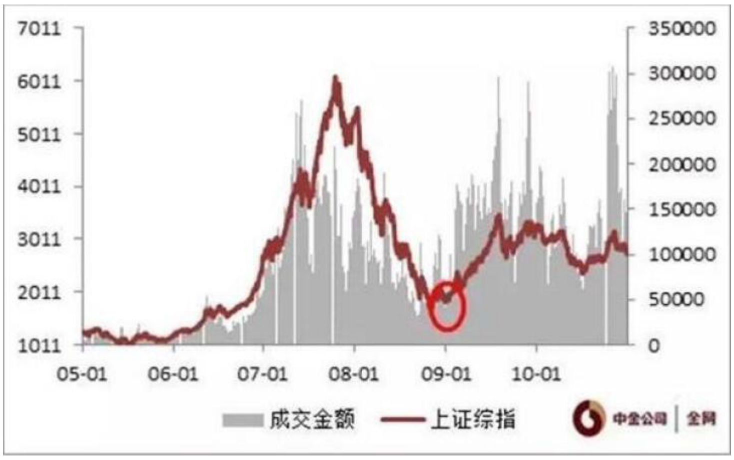

图 5　2005 年 1 月 -2010 年 12 月上证综指

应用过程

(1) 原始数据及处理：对 2005 年 1 月 -2016 年 3 月的上证综指时间序列作图，如图**??**，对序列做自相关图和偏自相关图分析，两者均为拖尾，说明序列不平稳，对序列进行一阶差分，如图6, 图中显示差分序列基本均匀分布在刻度上下两侧，因此可以认为差分序列平稳。

(2) 确定模型阶数：作差分序列的自相关图和偏自相关图，如图7。由图可知，差分序列的 ACF 和 PACF 都是拖尾的，对原序列建立 ARIMA(p,1,q) 模型，经过参数估计确定为 ARIMA(4,1,4) 模型。

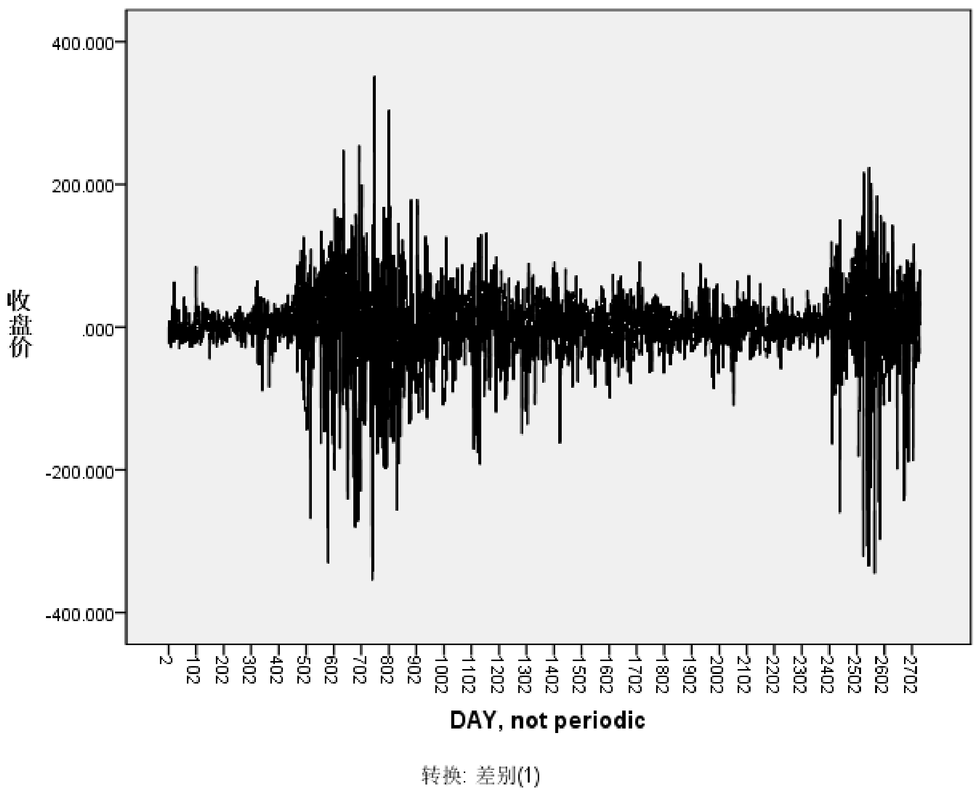

图 6　数据一阶差分图

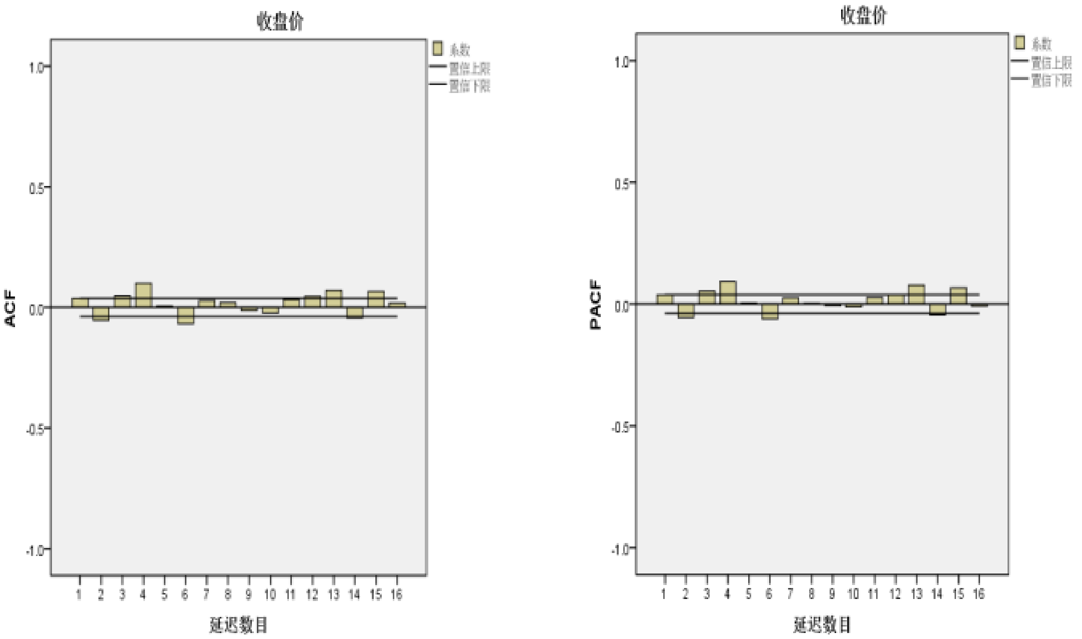

图 7　差分序列的 ACF（左）和 PACF（右）

(3) 残差序列检验：对 ARIMA(4,1,4) 模型进行残差序列是否为白噪声的检验，得到图8。由图认为残差序列为白噪声，ARIMA(4,1,4) 模型通过检验。

(4) 模型预测：使用 ARIMA(4,1,4) 模型进行预测，得到的结果如图9。

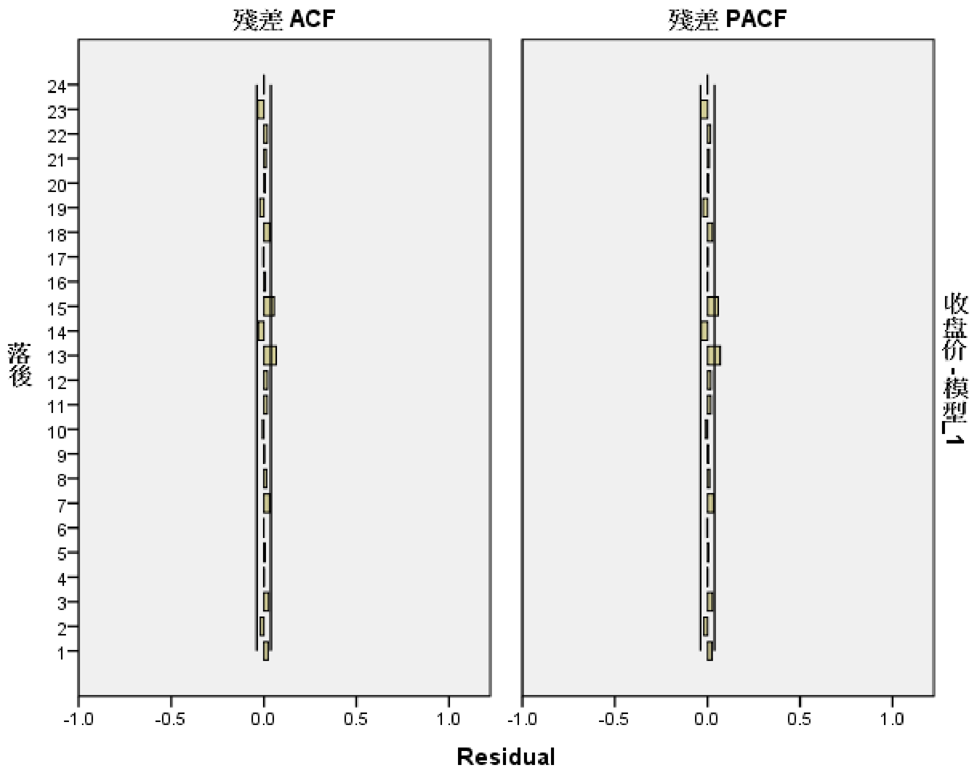

图 8　残差序列的 ACF（左）和 PACF（右）

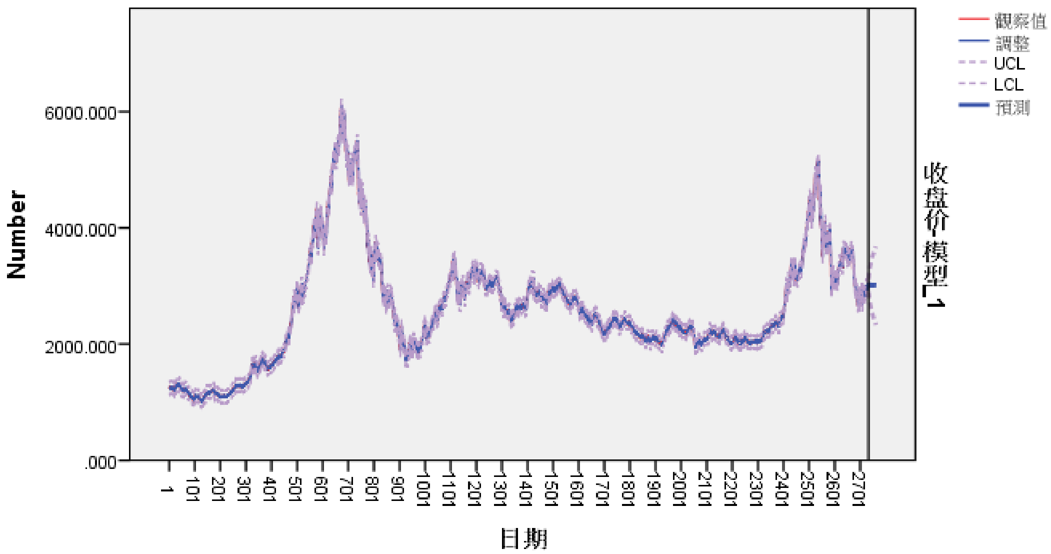

图 9　预测结果

四、参考文献

[1] Arima 模型. http://wiki.mbalib.com/wiki/ARIMA%E6%A8%A1%E5%9E%8B.

[2] Autoregressive integrated moving average. https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average.

[3] Arma 模型. `https://zh.wikipedia.org/wiki/ARMA%E6%A8%A1%E5%9E%8B`.

[4] 王燕. 应用时间序列分析. 中国人民大学出版社, 2005.

[5] 贾治宇，康锐. 基于 arima 模型的备件消耗预测方法. 兵工自动化, (28):29–31, 2009.

[6] 袁宇超、於江池、陈宏宇. 基于时间序列和神经网络的上证指数走势预测模型. 2016.