Advances in Applied Mathematics 应用数学进展, 2020, 9(7), 1103-1114
Published Online July 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.97130

Hans汉斯

Application of Time Series Analysis in Tea Yield Prediction in Fujian Province

Peiming Zhuang, Qitong Ou

School of Applied Mathematics, Xiamen University of Technology, XMUT, Xiamen Fujian
Email: 490053896@qq.com, ouqitong@xmut.edu.cn

Received: Jul. 6th, 2020; accepted: Jul. 22nd, 2020; published: Jul. 31st, 2020

Abstract

Based on the time series analysis theory and the SAS software, the total output of tea in Fujian Province from 1986 to 2015 was processed and analyzed. After the data processing, the time series was fitted with ARIMA(1,1,0) model and ARIMA(1,2,1) model. Model test and parameter test are carried out for the model respectively. After the test is passed, the two models are compared with the actual tea production of Fujian Province in 2016~2019, and the model with better fitting is determined by combining AIC criterion. Finally, using SAS software, we choose ARIMA(1,2,1) model to effectively predict the total output of tea in Fujian Province from 2020 to 2025.

Keywords

Time Series Analysis, SAS Software, ARIMA Model, Yield Prediction

时间序列分析在福建省茶叶产量预测中的应用

庄培铭，欧启通

厦门理工学院应用数学学院，福建 厦门
Email: 490053896@qq.com, ouqitong@xmut.edu.cn

收稿日期：2020年7月6日；录用日期：2020年7月22日；发布日期：2020年7月31日

摘　要

利用时间序列分析理论，通过使用SAS软件，对获取的福建省1986~2015年茶叶总产量进行数据处理及分析，通过数据处理后对构成的时间序列进行ARIMA(1,1,0)模型和ARIMA(1,2,1)模型的拟合。分别对模

文章引用: 庄培铭, 欧启通. 时间序列分析在福建省茶叶产量预测中的应用[J]. 应用数学进展, 2020, 9(7): 1103-1114.
DOI: 10.12677/aam.2020.97130

型进行模型检验以及参数检验，检验通过之后，分别用两个模型与福建省2016~2019年实际茶叶总产量进行检验，并结合AIC准则得出拟合较优的是ARIMA(1,2,1)模型。最后利用SAS软件，选择较优的ARIMA(1,2,1)模型对2020~2025年福建省茶叶总产量进行有效预测。

关键词

时间序列分析，SAS软件，ARIMA模型，产量预测

1. 引言

自改革开放以来，随着经济的发展，福建茶业得到快速发展。福建省气候地理环境十分适宜生产茶叶[1]，山地居多，气候温暖湿润，是我国茶叶的重要产地，有着一千多年的茶叶历史。茶业是福建九大支柱产业之一，对经济发展、农业增效和地区旅游业中起着不可替代的作用[2]。本文围绕着收集的福建省历年茶叶产量相关数据，进行序列平稳性判断及处理，然后对序列的自(偏)相关系数图进行研究，建立两个 ARIMA 时间序列模型，再检验分析两个模型的显著性以及参数显著性，根据近年福建省茶叶产量检验对比分析，选择预测效果较好的模型，来对福建省 2020~2025 年茶叶产量进行较高精度的拟合预测。

2. 茶叶产量数据的预处理

2.1. 平稳性判断

从福建省统计局官网收集到福建省 1986~2019 年的茶叶产量，如表 1 所示(单位：万吨)：

Table 1. Table of tea production in Fujian Province in 1986~2019
表 1. 福建省 1986~2019 年茶叶产量表

年份	茶叶产量	年份	茶叶产量	年份	茶叶产量	年份	茶叶产量
1986	4.42	1995	9.45	2004	16.44	2013	31.57
1987	4.99	1996	10.18	2005	18.48	2014	33.4
1988	5.54	1997	10.99	2006	20.01	2015	35.63
1989	5.52	1998	11.89	2007	22.09	2016	37.29
1990	5.82	1999	12.35	2008	24.07	2017	39.49
1991	6.53	2000	12.6	2009	25.51	2018	41.83
1992	7.05	2001	13.39	2010	25.83	2019	43.99
1993	7.7	2002	14.33	2011	27.67		
1994	8.24	2003	15.02	2012	29.6		

从中挑取 2016~2019 年的茶叶产量数据用来作为检验数据。

根据茶叶产量表的数据，利用 SAS 软件[3]，绘制福建省 1986~2015 年茶叶产量的时序图，如图 1 所示：

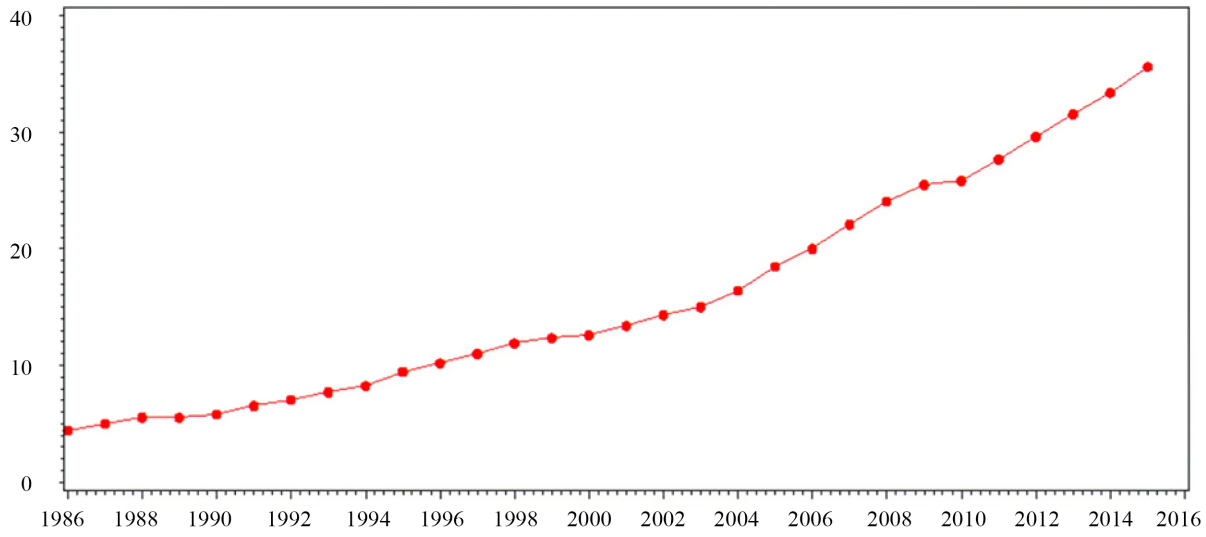

Figure 1. Time sequence of total tea production in Fujian Province from 1986 to 2015
图 1. 福建省 1986~2015 年茶叶总产量的时序图

Cramer 分解定理里说明了任何一个序列的波动都可以视为同时受到了确定性影响和随机影响的综合作用[4]。平稳序列要求这两方面的影响都是稳定的，而非平稳序列产生的机理则在于它所受到的这两方面的影响至少有一方面是不平稳的。

根据图 1 可以看出，时序图有明显的递增趋势，而非在一个常数值上下波动，可以判断出该序列为非平稳序列。

2.2. 平稳化处理

若要根据序列构建拟合模型，需要先进行差分运算，来通过自回归的方式提取确定性信息，使非平稳序列变为平稳[5]。运用 SAS 软件对该时间序列进行一阶差分，并画出时序图，如图 2 所示：

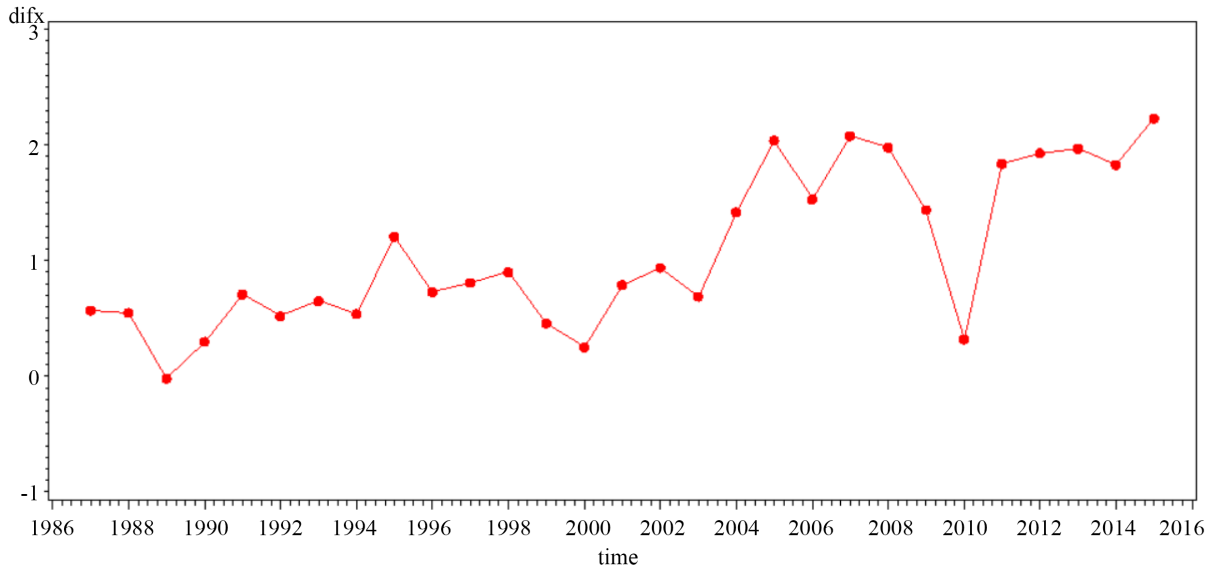

Figure 2. Time sequence diagram of the first-order difference of the total output of tea in Fujian Province from 1986 to 2015
图 2. 福建省 1986~2015 年茶叶总产量一阶差分后的时序图

做一阶差分后，时序图没有明显的递增或者递减趋势，也没有明显周期性的变化，为了准确判断差分后的平稳性，运用 Eviews 软件[6]来对一阶差分后的序列做 ADF 检验。检验结果如图 3 所示：

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-3.934622	0.0238
Test critical values:	1% level	-4.323979	
	5% level	-3.580623	
	10% level	-3.225334	

Figure 3. Unit root test after first order difference of sequence
图 3. 序列一阶差分后的单位根检验

检验结果显示，P 值为 0.0238 小于显著性水平 0.05，拒绝存在单位根的原假设，所以一阶差分后的序列是平稳的。

2.3. 纯随机性检验

纯随机性检验也称为白噪声检验，用来检验一个序列是否为纯随机序列。若一个序列为纯随机序列，那么该时间序列的序列值之间没有任何相关关系，这就意味着序列值之间不会互相影响，也就没有继续研究下去的价值。根据 Bartlett 定理[7]，Ljung 和 Box 构造了 LB 统计量来进行检验：

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

其中，n 为序列观察期数，m 为指定延迟期数。

运用 SAS 软件计算出各个延迟阶数下的 LB 检验统计值的量和 P 值，结果如表 2 所示：

Table 2. White noise test chart
表 2. 纯随机性检验表

纯随机性检验		
延迟阶数	LB 检验统计量的值	P 值
6	31.14	<0.0001
12	38.98	0.0001

检验结果显示，各阶延迟阶数下的 LB 统计量的 P 值都显著小于显著性水平 0.05，拒绝为白噪声序列的原假设，所以该序列为平稳非白噪声序列。

3. ARIMA(1,1,0)模型识别及参数估计

3.1. 模型识别

福建省 1986~2015 年茶叶总产量一阶差分后的自相关图和偏自相关图如图 4 和图 5 所示。通过观察自相关图和偏自相关图可以看出：当延迟期数大于 1 之后，自相关系数都在两倍标准差之内，且呈现指数衰减到零附近，呈现出拖尾性质，所以自相关系数为一阶拖尾；当延迟阶数为 1 时，偏自相关系数明显大于 2 倍标准差范围，当延迟阶数大于 1 后，迅速衰减到两倍标准差之内，且几乎都落在 2 倍标准差范围以内，呈现截尾性质，所以偏自相关系数为一阶截尾。所以把该模型定阶为 ARIMA(1,1,0)模型[8]。

Autocorrelations

Lag	Covatiance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.434272	1.00000	|********************|	0
1	0.277512	0.63903	|*************.	0.185695
2	0.203445	0.46847	|*********.	0.250290
3	0.169127	0.38945	|********.	0.278892
4	0.128907	0.29683	|******.	0.297054
5	0.032710	0.07532	|**.	0.307112
6	0.100493	0.23141	|*****.	0.307748
7	0.128260	0.29535	|******.	0.313691
8	0.065857	0.15165	|***.	0.323137
9	0.055547	0.12791	|***.	0.325582
10	0.080764	0.18598	|****.	0.327310
11	0.0043289	0.0997	|.	0.330934
12	-0.057244	-0.13182	.***|	0.330944
13	-0.067292	-0.15495	.***|	0.332750
14	-0.092821	-0.21374	.****|	0.335229
15	-0.149163	-0.34348	.*******|	0.339896

Figure 4. Autocorrelation graph after first order difference of sequence
图 4. 序列一阶差分后的自相关图

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
1	0.63903	|*************.
2	0.10161	|**.
3	0.09490	|**.
4	-0.00985	|.
5	-0.26165	.*****|
6	0.41744	|********.
7	0.09679	|**.
8	-0.26424	.*****|
9	0.08457	|**.
10	-0.01156	|.
11	-0.16470	.***|
12	-0.06257	.*|
13	-0.26333	.*****|
14	-0.03574	.*|
15	-0.02728	.*|

Figure 5. Partial autocorrelation graph after first order difference of sequence
图 5. 序列一阶差分后的偏自相关图

3.2. 模型参数估计

模型定阶后，对模型未知的参数值进行估计。对于时间序列中的未知参数的估计方法有三种：矩估计、极大似然估计、最小二乘估计[7]。在实际运用中，最小二乘估计法是最常用的方法。最小二乘估计法在 ARMA 模型场合中，记

$$\tilde{\beta}=\left(\varphi_1,\cdots,\varphi_p,\theta_1,\cdots,\theta_q\right)' \tag{2}$$

$$F_t\left(\tilde{\beta}\right)=\varphi_1 x_{t-1}+\cdots+\varphi_p x_{t-p}-\theta_1\varepsilon_{t-1}-\cdots-\theta_q\varepsilon_{t-q} \tag{3}$$

当它假定过去未观测到的序列值等于 0，那么可以得到它的残差平方和为：

$$Q\left(\tilde{\beta}\right)=\sum_{t=1}^{n}\left[x_t-\sum_{i=1}^{t}\pi_i x_{t-i}\right]^2 \tag{4}$$

使此式达到最小值的估计值就为参数 β 的最小二乘估计。通过 SAS 软件，运用条件最小二乘估计法来对拟合的模型进行参数估计，结果如表 3 所示：

Table 3. Parameter estimation table
表 3. 参数估计表

条件最小二乘估计			
参数	估计值	标准误差	滞后
MU	1.03544	0.28087	0
AR1,1	0.71939	0.15027	1

由表 3 可得，拟合的模型为：

$$\nabla x_t = 1.03544 + \frac{\varepsilon_t}{1-0.71939B} \tag{5}$$

或等价表示为

$$x_t = 0.29056 + 1.71939x_{t-1} - 0.71939x_{t-2} + \varepsilon_t \tag{6}$$

4. ARIMA(1,1,0)模型检验及预测

4.1. 模型检验

确定拟合模型的口径后，对模型进行显著性检验和参数显著性检验。显著性检验即通过检验残差序列是否纯随机来检验模型的信息是否提取充分，以来判定模型的估计效果；参数性检验即检验模型各参数是否显著非零，来评估模型的可行性和精简性，不显著非零的参数则表明该参数对应的自变量对因变量影响不明显，需要从拟合的模型中剔除。通过 SAS 软件获取模型显著性检验表以及参数显著性检验表，结果如表 4、表 5 所示：

Table 4. Model significance test table
表 4. 模型显著性检验表

纯随机性检验		
延迟阶数	LB 检验统计量的值	P 值
6	7.20	0.2062
12	16.27	0.1314
18	22.79	0.1561
24	28.77	0.1882

显著性检验表结果显示，各阶延迟阶数下的 LB 统计量的 P 值都显著大于 0.05，所以认为这个拟合模型的残差序列属于白噪声序列，即该模型显著有效。

Table 5. Parameter significance test table
表 5. 参数显著性检验表

参数	t 统计量的值	P 值	结论
μ	3.69	0.001	显著非零
φ_1	4.79	<0.0001	显著非零

参数显著性检验表结果显示，所有参数均通过显著非零检验，说明 ARIMA(1,1,0)模型对该序列的拟合显著成立。

4.2. 模型预测及数据对比

根据拟合的模型公式，利用 SAS 软件对福建省 2016~2019 年的茶叶总产量进行预测，结果如表 6 所示：

Table 6. Prediction table of total output value of tea in Fujian Province in 2016~2019
表 6. 2016~2019 年福建省茶叶总产量值预测表

年份	预测值/万吨	标准差	95%置信区间
2016	37.5248	0.5033	(36.5384,38.5111)
2017	39.1784	1.0010	(37.2165,41.1403)
2018	40.6586	1.5064	(37.7061,43.6111)
2019	42.0140	1.9984	(38.0973,45.9307)

把福建省 2016~2019 年茶叶产量数据摘出，用来与预测值作对比，来判断模型预测是否较为成功。对比数据如表 7 所示：

Table 7. Comparison table of total output value of tea in Fujian Province in 2016~2019
表 7. 2016~2019 年福建省茶叶总产量值对比表

年份	预测值/万吨	真实值/万吨	绝对误差	误差率%
2016	37.5248	37.29	0.2348	0.6
2017	39.1784	39.49	0.3116	0.7
2018	40.6586	41.83	1.1714	2.7
2019	42.01	43.99	1.98	4.5

从表 7 中可以看出，模型总体的预测精度良好，误差率总体较低，能基本稳定在 1%以内。但随着预测步的增加，实际数据与预测数据的误差率逐渐增大。

5. ARIMA(1,2,1)模型识别及参数估计

5.1. 平稳性检验

运用 SAS 软件，将原始序列进行二阶差分，得到的时序图如图 6 所示：

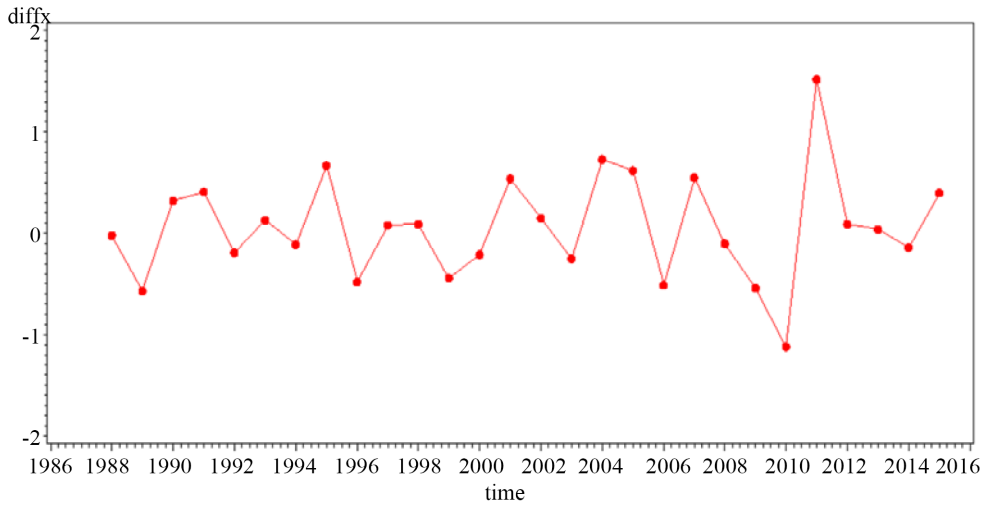

Figure 6. Sequence diagram after second order difference
图 6. 序列二阶差分后的时序图

二阶差分后的时序图没有明显的趋势走向，也没有周期性的变化，为了准确判断平稳性，对序列进行 ADF 检验。检验结果如图 7 所示。结果显示，P 值为 0 小于 0.05，拒绝存在单位根的原假设，所以可以判断二阶差分后的序列为平稳序列。

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-6.612957	0.0000
Test critical values:	1% level	-2.653401	
	5% level	-1.953858	
	10% level	-1.609571	

Figure 7. Unit root test after second order difference of sequence
图 7. 序列二阶差分后的单位根检验

5.2. 纯随机性检验

检验序列为平稳序列之后，对序列进行纯随机性检验。利用 SAS 软件得到二阶差分后序列的纯随机性检验表，自相关图和偏自相关图。纯随机性检验表如表 8 所示，检验结果显示，各阶延迟阶数下的 LB 统计量的 P 值都小于 0.05，拒绝原假设，所以该序列为平稳非白噪声序列。

Table 8. Second order difference pure random test table
表 8. 二阶差分纯随机检验表

纯随机性检验		
延迟阶数	LB 检验统计量的值	P 值
6	12.75	0.0472
12	22.39	0.0334

6. ARIMA(1,2,1)模型识别及参数估计

6.1. 模型识别

序列二阶差分后的自相关图和偏自相关图如图 8、图 9 所示，通过观察自相关图和偏自相关图可以看出：自相关系数快速地进入两倍标准差，且延迟一阶之后的自相关系数都在二倍标准差之内，所以自相关系数为一阶截尾。当延迟阶数大于等于 1 时，偏自相关系数都在两倍标准差之内，所以可以判断该时间序列的偏自相关系数为一阶截尾。所以把该模型定阶为 ARIMA(1,2,1)模型。

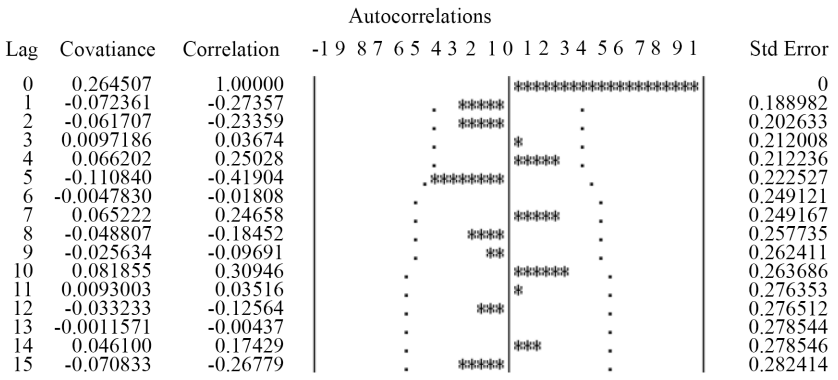

Figure 8. Autocorrelation graph after second order difference of sequence
图 8. 序列二阶差分后的自相关图

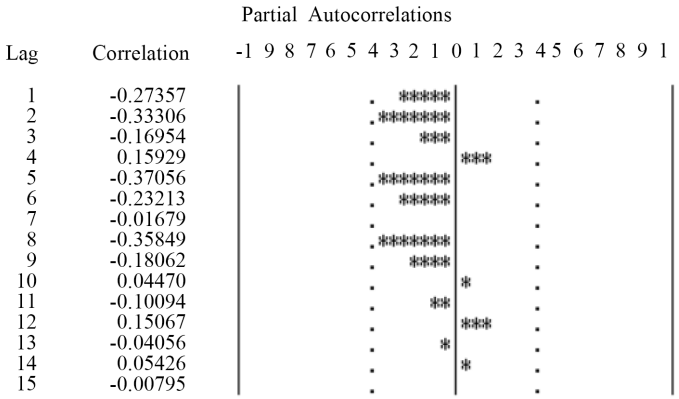

Figure 9. Partial autocorrelation graph after second order difference of sequence
图 9. 序列二阶差分后的偏自相关图

6.2. 模型参数估计

将二阶差分后的序列模型定阶后，通过 SAS 软件，运用条件最小二乘估计法来对拟合的模型进行参数估计，结果如表 9 所示：

Table 9. Parameter estimation table after second order difference
表 9. 二阶差分后的参数估计表

条件最小二乘估计			
参数	估计值	标准误差	滞后
MU	0.04228	0.03276	0
MA1,1	1	0.19726	1
AR1,1	0.47205	0.26403	1

由得到的参数估计表可知，拟合的模型为：

$$\nabla^2 x_t = 0.04228 + \frac{1-B}{1-0.47205B}\varepsilon_t \tag{7}$$

7. ARIMA(1,2,1)模型检验及预测

确定拟合模型的口径后，对模型进行显著性检验和参数显著性检验，检验结果如表 10、表 11 所示：

Table 10. Significance test table of second-order difference model
表 10. 二阶差分后模型显著性检验表

纯随机性检验		
延迟阶数	LB 检验统计量的值	P 值
6	8.26	0.0825
12	17.01	0.0742
18	22.92	0.1160
24	29.39	0.1341

检验结果显示，各阶延迟下的 LB 统计量的 P 值都显著大于显著性水平 0.05，所以可以认为这个拟

合模型的残差序列属于白噪声序列，即该模型显著有效。

Table 11. Test table of parameter significance after second-order difference
表 11. 二阶差分后参数显著性检验表

参数	t 统计量的值	P 值	结论
μ	1.29	0.2087	不显著非零
θ_1	5.07	<0.0001	显著非零
φ_1	1.79	0.0859	不显著非零

检验结果显示，参数显著性检验的常数项 μ 和参数值 φ_1 未通过显著非零检验，所以这两个参数值为零。把两个不显著非零的参数剔除后，得到了新的参数显著性检验表(表 12)：

Table 12. Parameter significance test table after parameter elimination
表 12. 参数剔除后参数显著性检验表

参数	t 统计量的值	P 值	结论
θ_1	2.89	0.0075	显著非零

剔除掉不显著非零的参数之后，再通过 SAS 软件，运用条件最小二乘估计法来对拟合的模型进行参数估计，结果如表 13 所示：

Table 13. Parameter estimation table after parameter elimination
表 13. 参数剔除后的参数估计表

条件最小二乘估计			
参数	估计值	标准误差	滞后
MA1,1	0.49526	0.17118	1

由二阶差分参数剔除后的参数估计表所知，拟合的模型为：

$$x_t = 2x_{t-1} - x_{t-2} + 0.50424\varepsilon_t \tag{8}$$

根据二阶差分后拟合的模型公式，利用 SAS 软件对福建省 2016 年~2019 年的茶叶总产量进行预测，结果如表 14 所示：

Table 14. Prediction table of total output value of tea in Fujian Province in 2016~2019
表 14. 2016~2019 年福建省茶叶总产量值预测表

年份	预测值/万吨	标准差	95%置信区间
2016	37.6606	0.4909	(36.6985,38.6226)
2017	39.6911	0.8868	(37.9529,41.4293)
2018	41.7217	1.3264	(39.1219,44.3214)
2019	43.7522	1.8117	(40.2013,47.3032)

和一阶差分一样，把 2016~2019 的数据摘出，做预测之后与预测值作对比，来判断模型拟合是否较为成功。对比数据如表 15 所示：

Table 15. Comparison table of total output value of tea in Fujian Province in 2016~2019
表 15. 2016~2019 年福建省茶叶总产量值对比表

年份	预测值/万吨	真实值/万吨	绝对误差	误差率%
2016	37.6606	37.29	0.3706	0.9
2017	39.6911	39.49	0.2011	0.5
2018	41.7217	41.83	0.1083	0.4
2019	43.7522	43.99	0.2378	0.5

从表 15 中可以看出，模型的预测值与真实值的误差稳定在 0.5 内，误差率总体较低，稳定在 1%以内，整体的误差比 ARIMA(1,1,0)模型的还要小，说明模型 ARIMA(1,2,1)的总体预测精度更好。

8. 模型对比及预测

模型结果对比如表 16 所示，可以看出 ARIMA(1,2,1)模型相比 ARIMA(1,1,0)模型的福建省 2016~2019 年茶叶产量预测值，更接近数据的真实值，且 ARIMA(1,2,1)模型的 AIC 值[9]小于 ARIMA(1,1,0)模型，根据 AIC 准则，ARIMA(1,2,1)模型会优于 ARIMA(1,1,0)模型。

Table 16. Comparison table of model results
表 16. 模型结果对比表

年份	ARIMA(1,1,0)模型	ARIMA(1,2,1)模型	真实值
2016	37.5248	37.6606	37.29
2017	39.1784	39.6911	39.49
2018	40.6586	41.7217	41.83
2019	42.01	43.7522	43.99
AIC 值	44.39	40.59	

结合多种因素进行结合考虑，最终选择拟合较优的 ARIMA(1,2,1)模型对未来几年福建省茶叶产量进行预测，结果如表 17：

Table 17. Prediction table of total output value of tea in Fujian Province from 2020 to 2025
表 17. 2020~2025 年福建省茶叶总产量值预测表

年份	预测产量/万吨	标准差	95%置信区间
2020	45.7828	2.3406	(41.1953,50.3702)
2021	47.8133	2.9103	(42.1092,53.5174)
2022	49.8439	3.5185	(42.9477,56.7400)
2023	51.8744	4.1631	(43.7150,60.0339)
2024	53.9050	4.8421	(44.4146,63.3953)
2025	53.9355	5.5541	(45.0497,66.8213)

根据预测的结果，福建省未来的茶叶产量依旧呈上升趋势，也有可能逐渐趋于平稳。

9. 结论

福建省一直致力茶叶品牌建设与开发，茶业又好又快发展的成因，除了自身环境条件之外，主要还

有行业相关和政府部门的重视，予以政策扶持，以及茶叶科技先进技术应用的进步。本文利用 SAS 软件，运用时间序列分析的理论知识，对福建省的茶叶总产量进行研究，通过建立的两个模型预测福建省茶叶总产量未来趋势，根据数据比对以及各因素综合考虑挑选出拟合较好的一个模型。通过该模型体现的茶叶产量发展，为有效的未来茶业发展提供参考依据。

参考文献

[1] 沈德福. 福建三大品牌发展现状与开发路径研究[J]. 福建茶叶, 2019(8): 1-2.

[2] 沈德福. 福建省近二十年茶叶发展态势及地区资源禀赋差异分析[J]. 茶叶, 2019(1): 50-52.

[3] 肖枝洪, 郭明月. 时间序列分析与 SAS 应用[M]. 武汉: 武汉大学出版社, 2009.

[4] 刘佳, 赵慧文, 刘光荣. 基于 SAS 的非平稳时间序列分析及实证研究[J]. 汕头大学学报·自然科学, 2010, 25(1): 48-53.

[5] 张亚婕. 基于 ARIMA 模型对股票和指数预测结果的简单比较分析[J]. 市场研究, 2019(11): 23-25.

[6] 于俊年. 计量经济学软件 Eviews 的使用[M]. 北京: 对外经济贸易大学出版社, 2012.

[7] 易丹辉, 王燕. 应用时间序列分析[M]. 北京: 中国人民大学出版社, 2019.

[8] 阮敬, 纪宏编. 使用 SAS 统计分析教程[M]. 北京: 中国统计出版社, 2013.

[9] 郑岩岩. 基于 SAS 软件的时间序列分析在 GDP 预测中的应用[J]. 金融经济, 2013(9): 179-181.