Statistics and Application 统计学与应用, 2021, 10(1), 162-172
Published Online February 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.101016

Hans汉斯

基于季节效应ARIMA模型对贵州省城镇新增就业人数的预测

肖　霏

曲阜师范大学统计学院，山东　济宁
Email: x_feidem@163.com

收稿日期：2021年1月25日；录用日期：2021年2月19日；发布日期：2021年2月26日

摘　要

“脱贫攻坚”是一场长时间的战斗，脱贫攻坚重点工作扎实推进，其中贫困县较多的贵州省扶贫成果尤为显著。选取2002年至2019年贵州省城镇就业增长人数的月度数据进行时间序列分析，建立季节乘积的ARIMA模型，并运用所建模型对2020年贵州省城镇就业增长人数进行预测，进行稳健性检验以说明预测的可靠性。预测结果表明2020年城镇就业新增人口仍将保持长期增长趋势。

关键词

时间序列分析，ARIMA模型，脱贫攻坚，城镇新增就业人数

Prediction of Newly-Increased Employment in Urban Areas of Guizhou Province Based on Seasonal Effect ARIMA

Fei Xiao

School of Statistics, Qufu Normal University, Jining Shandong
Email: x_feidem@163.com

Received: Jan. 25th, 2021; accepted: Feb. 19th, 2021; published: Feb. 26th, 2021

Abstract

“Poverty alleviation” is a long-term battle, and the key tasks of poverty alleviation have been steadily advanced. Among them, Guizhou Province, which has more impoverished counties, has

文章引用：肖霏. 基于季节效应 ARIMA 模型对贵州省城镇新增就业人数的预测[J]. 统计学与应用, 2021, 10(1): 162-172. DOI: 10.12677/sa.2021.101016

achieved remarkable results in poverty alleviation. This paper selects monthly data of the number of urban employment growth in Guizhou Province from 2002 to 2019 for time series analysis, establishes a seasonal product ARIMA model, and uses the built model to predict the number of urban employment growth in Guizhou Province in 2020, conducting a robustness test to illustrate reliability of predictions. The forecast results indicate that the newly-increased urban employment population in 2020 will still maintain seasonality and growth.

Keywords

Time Series Analysis, ARIMA Model, Alleviate Poverty, New Employment in Urban Areas

1. 引言

贵州省脱贫攻坚消除贫困、改善民生、逐步实现共同富裕，是社会主义的本质要求，也是中国共产党的重要使命。本文依据幸福指标评价构建的三大原则，包括区域性、代表性及可操作性[1]，针对贵州省的实际情况，脱贫攻坚战略实施以来所取得的成就[2]，从城镇就业增长人数角度分析近几年来贵州省发展情况。刘攀结合贵州林业推进扶贫攻坚优势和采取的举措，提出做好林业扶贫攻坚工作的对策建议[3]。于冰认为每个阶段的不同情况，导致脱贫攻坚存在的问题也在不断变化，要精准识别发展转向和矛盾所在[4]。

贵州省在保障城镇居民就业方面的工作稳步开展，绝大多数的城镇居民基本生活得到保障。冉茂文研究发现，城镇发展能够辐射周围经济欠发达的地方，贵州省脱贫重心应从农村转向城市[5]。李如是基于灰色关联度分析指出城镇新增就业人口变化具有一定周期性[6]。吴江认为城镇新增就业人数受到当地经济发展，产业结构，教育水平等多方面的影响[7]。新增就业人数能够较为直观地反映贵州省脱贫攻坚的进展，因此对新增就业人数未来走势的预测，能为未来脱贫工作的研究提供参考。预测是决策的基础和依据，对新增就业人数进行预测，能够为相关政策措施的实施提供一定的帮助，所以对贵州省新增就业人数的预测是有意义且有必要的。

2. 理论基础

2.1. 季节乘积 ARIMA 模型

ARIMA 模型(Autoregressive Integrated Moving Average model)，差分整合移动平均自回归模型，又称整合移动平均自回归模型，是时间序列预测分析方法之一[8]。ARIMA(p,d,q)中，AR 是“自回归”，p 为自回归项数；MA 为“滑动平均”，q 为滑动平均项数，d 为使之成为平稳序列所做的差分次数。而季节乘积 ARIMA 模型就是 ARIMA 模型和季节模型的综合。如果时间序列 Y_t 除了趋势变动外，还有较明显的季节性变动，就先要对序列进行逐期差分，消除趋势性，再进行季节差分消除序列的季节性，差分步长应与季节周期一致，然后建立包含季节差分的有关参数的 ARIMA 模型，即季节乘积 ARIMA 模型。ARIMA(p,d,q) × $(P,D,Q)_S$ 模型的一般表达式为：

$$\varphi_p(L)\Phi_P\left(L^s\right)(1-L)^d\left(1-L^s\right)^D Y_t = \theta_q(L)\theta_Q\left(L^s\right)\varepsilon_t$$

其中，$\varphi_p(L)=1-\varphi_1 L-\varphi_2 L^2-\cdots-\varphi_p L^p$ 是非季节自回归多项式，p 是自回归阶数；

$\theta_p(L)=1-\theta_1 L-\theta_2 L^2-\cdots-\theta_q L^q$ 是非季节移动平均多项式，q 是移动平均阶数；

$\theta_p(L)=1-\theta_1 L-\theta_2 L^2-\cdots-\theta_q L^q$ 是季节自回归多项式，其中 P 是季节自回归阶数；

$\theta_Q(L)=1-\theta_1 L-\theta_2 L^2-\cdots-\theta_Q L^Q$ 是季节移动平均多项式，其中 Q 是季节移动平均阶数；

$(1-L)^d$ 为差分算子，d 为差分阶数；$\left(1-L^s\right)^D$ 为季节差分算子，D 为季节差分阶数，s 为季节周期。

2.2. 确定性因素分解方法

确定性因素分解方法的基本思想是：尽管不同的序列波动特征千差万别，但是序列的各种变化都可以归纳成四大类因素的综合影响[8]：

1) 长期趋势(trend)：该因素的影响会导致序列呈现出明显的长期趋势(递增、递减等)；

2) 循环波动(circle)：该因素会导致序列呈现出从低到高再由高至低的反复循环波动，如果观察时期不够长则改为交易日(day)因素；

3) 季节性变化(season)：该因素会导致序列呈现出和季节变化相关的稳定的周期波动；

4) 随机波动(immediate)：除了长期趋势、循环波动(交易日)、季节性变化之外，序列还会受到各种其他因素的综合影响，而这些影响导致序列呈现出一定的随机波动。

3. 分析与预测

3.1. 新增就业人口走势分析

贵州省 2002 年 1 月至 2019 年 12 月城镇新增就业人数的历史数据序列记作 Y，从时序图(图 1)可以大致看出其历史走势与波动状况，图中横坐标为年份，纵坐标表示新增就业人数。

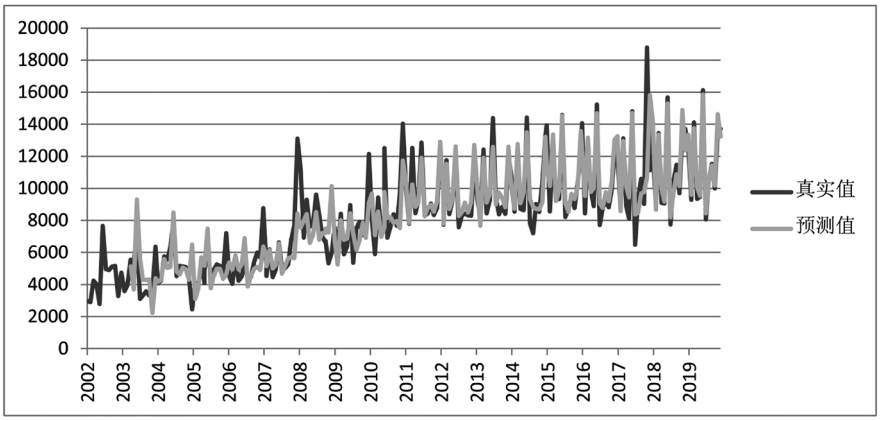


Figure 1. Time series of newly-increased urban employment population from 2002 to 2019
图 1. 2002 年至 2019 年城镇新增就业人口时序图

城镇新增就业人数呈现明显的增长趋势，且在一定程度上受到季节效应因素的影响。这里是广义的“季节”，即凡是呈现出固定的周期性变化的事件都可以称它具有“季节”效应。从图 2 可以看到新增就业人口数量存在明显的季节效应，3 月和 6 月新增人数明显高于其他月份。

总体来看，新增就业人口呈现出逐渐上升的趋势性以及明显的季节性特征，季节性波动的趋势性以及明显的季节性特征，季节性波动，速度逐渐放缓。

由于贵州省 2016 年 1 月至 2019 年 10 月城镇新增就业人数序列显示出一定的规律性，如该序列有显著的增长趋势、有固定的变化周期等，因此考虑采用确定性分析方法，这里采用常用的确定性因素分解方法。

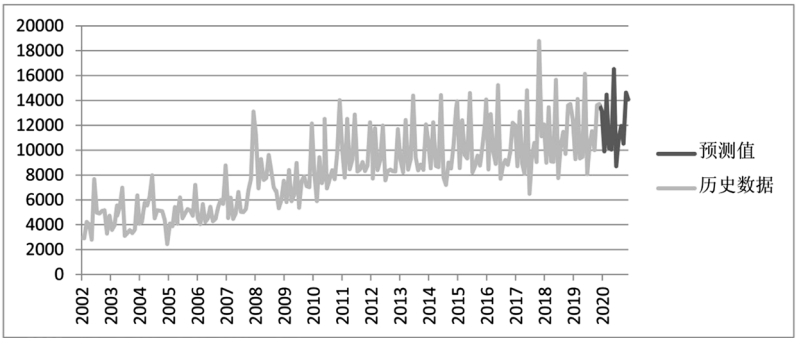


Figure 2. Seasonal fluctuations in new employment
图 2. 新增就业人口的季节性波动

由于指标的量级不同，当各指标间的水平相差很大时，若直接利用原始指标值进行分析，就会突出数值水平较高的指标在综合分析中的作用，降低数值水平较低指标的作用。为了保证结果的可靠性，在数据整理完成后，再进一步通过 SPSS 软件对数据进行标准化处理，处理步骤：SMS 数据窗口依次点击：分析－描述统计－描述，将各变量选入备选框，选择将标准化得分另存为变量，点击确定，在数据窗口得到各变量的标准化得分[9]。

3.2. 建立 ARIMA 模型

由于从预测角度看，近期的数值要比远期的数值对未来有更大作用，为了利用更多历史数据的同时使得模型预测的精确度更高，本文选取 2002 年 1 月至 2019 年 12 月共 18 年间新增就业人数的月度数据。其中 2002 年至 2018 年的所有数据用作建立 ARIMA 模型的样本，2019 年数据用于样本内预测以检验模型的拟合效果。

本文采用 R 3.6.3 软件和 Eviews10.0 软件建立模型并进行预测。ARIMA 模型的建立流程如图 3 所示。

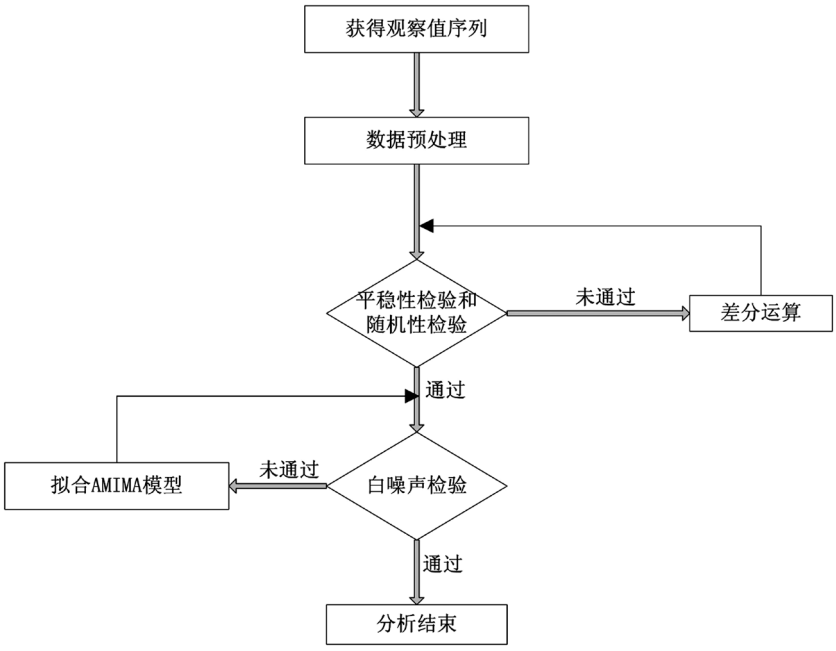


Figure 3. ARIMA modeling flowchart
图 3. ARIMA 建模流程图

建立 ARIMA 模型通常有时间序列的预处理、模型识别、模型定阶、参数估计、模型有效性验证几个步骤。

1) 时间序列数据的预处理

时间序列必须同时满足平稳性与非白噪声两个条件才能适用 ARIMA 模型进行分析预测。如果数据是非平稳的，可使用指数化或差分的方式将其变为平稳数据，若序列是白噪声的，说明各数值之间没有明显的相关关系，即过去的情况对未来的趋势发展没有影响，所以没有价值，因此对数据的预处理包括平稳性检验与白噪声检验两部分。首先可以从该序列的时序图来观察其平稳性，从图 1 可以看出新增就业人数有增长趋势，并且有明显的季节性波动，数据是非平稳的。用单位根检验看其平稳性(结果见表 1)，可以看到 ADF 检验统计量值为−2.5690，大于检验水平为 1%、5%、10%的临界值，所以存在单位根，该序列是非平稳的。

Table 1. Unit root test for sequence Y
表 1. 序列 Y 的单位根检验

Augmented Dickey-Fuller		t-Statistic	Prob
test statistic		−2.5690	0.2951
	1%	−4.0039	
Test critical values	5%	−3.4321	
	10%	−3.1398	

为使序列变平稳，将原序列 Y 进行一阶差分处理，得到序列 DY。

从图 4 可以看出差分后的序列 DY 始终在一个常数值上下波动，趋势性消除，对其进一步进行单位根检验(结果见表 2)，可知此时拒绝存在单位根的原假设，序列 DY 是平稳的。

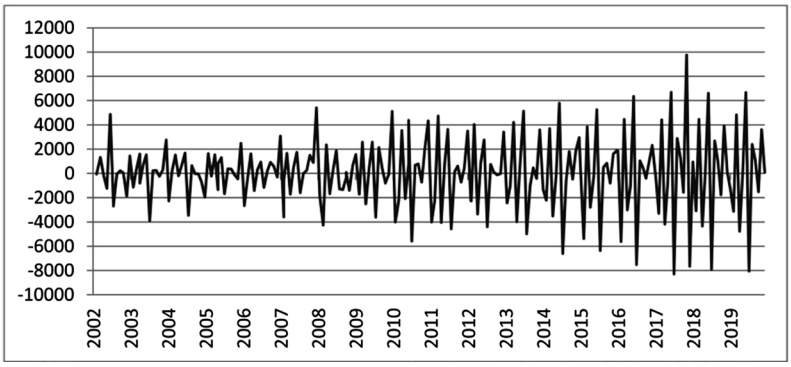
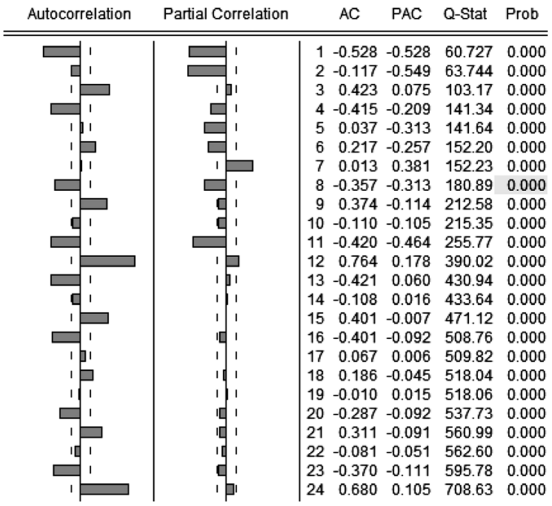


Figure 4. Timing diagram of sequence DY after differing
图 4. 差分后的序列 DY 的时序图

Table 2. Unit root test for sequence Y
表 2. 序列 Y 的单位根检验

Augmented Dickey-Fuller		t-Statistic	Prob
test statistic		−6.7022	0.2951
	1%	−2.5764	
Test critical values	5%	−1.9424	
	10%	−1.6157	

平稳性检验通过后，还要对序列 DY 进行白噪声检验，因此做出其自相关与偏自相关图进行进一步考察，从图 5 可看到，在 6 阶、12 阶、18 阶、24 阶，Q 统计量的 P 值均小于 0.05，因此该序列为非白噪声序列。通过上述检验，可知 DY 为平稳非白噪声序列，可以建立 ARIMA 模型。



	AC	PAC	Q-Stat	Prob
1	-0.528	-0.528	60.727	0.000
2	-0.117	-0.549	63.744	0.000
3	0.423	0.075	103.17	0.000
4	-0.415	-0.209	141.34	0.000
5	0.037	-0.313	141.64	0.000
6	0.217	-0.257	152.20	0.000
7	0.013	0.381	152.23	0.000
8	-0.357	-0.313	180.89	0.000
9	0.374	-0.114	212.58	0.000
10	-0.110	-0.105	215.35	0.000
11	-0.420	-0.464	255.77	0.000
12	0.764	0.178	390.02	0.000
13	-0.421	0.060	430.94	0.000
14	-0.108	0.016	433.64	0.000
15	0.401	-0.007	471.12	0.000
16	-0.401	-0.092	508.76	0.000
17	0.067	0.006	509.82	0.000
18	0.186	-0.045	518.04	0.000
19	-0.010	0.015	518.06	0.000
20	-0.287	-0.092	537.73	0.000
21	0.311	-0.091	560.99	0.000
22	-0.081	-0.051	562.60	0.000
23	-0.370	-0.111	595.78	0.000
24	0.680	0.105	708.63	0.000

Figure 5. Autocorrelation graph and partial autocorrelation graph of sequence DY
图 5. 序列 DY 的自相关图与偏自相关图

2) 模型识别

序列 DY 没有明显的拖尾和截尾特征，难以按照传统方式定阶。可以注意到自相关图中。延迟 12 阶和 24 阶时，自相关系数达到最大且显著大于两倍标准差，可以推测序列 DY 具有以 12 为周期的季节性波动特征，因此可以将序列 DY 进行季节调整。将序列 DY 进行十二步差分，得到一阶十二步差分序列 DY12，时序图、自相关及偏自相关图分别见图 6 和图 7。

从季节调整后的时序图(图 6)可以看到受季节因素影响的波动减小，但是自相关图和偏自相关图(图 7)截尾和拖尾趋势仍不明显，尝试拟合 ARIMA 模型但是效果较差。观察到在延迟 12 阶时 AIC 和 PAIC 仍较大，在图中仍明显突出，可以认为季节调整后的序列中仍存在季节效应，因此该序列的短期相关性与季节效应有复杂的关联性，不能简单提取，可以尝试拟合乘积季节 ARIMA$(p,d,q) \times (P,D,Q)_S$ 模型。

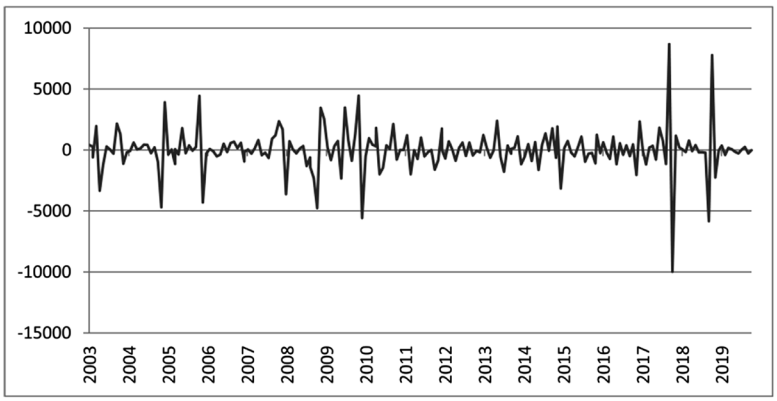
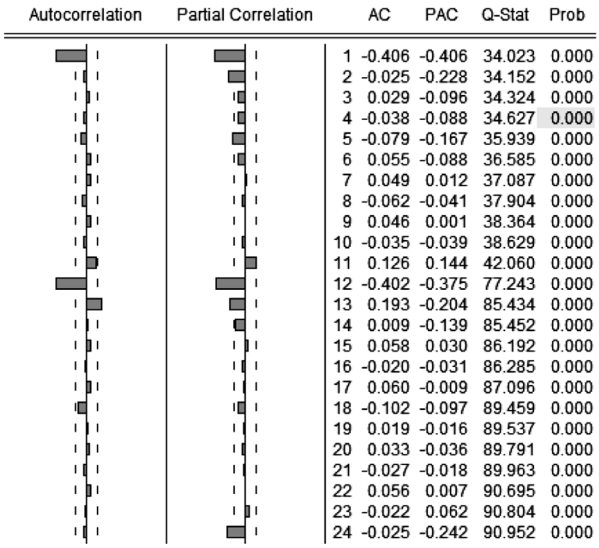


Figure 6. Timing diagram of seasonally adjusted sequence DY12
图 6. 季节调整后序列 DY12 的时序图



	AC	PAC	Q-Stat	Prob
1	-0.406	-0.406	34.023	0.000
2	-0.025	-0.228	34.152	0.000
3	0.029	-0.096	34.324	0.000
4	-0.038	-0.088	34.627	0.000
5	-0.079	-0.167	35.939	0.000
6	0.055	-0.088	36.585	0.000
7	0.049	0.012	37.087	0.000
8	-0.062	-0.041	37.904	0.000
9	0.046	0.001	38.364	0.000
10	-0.035	-0.039	38.629	0.000
11	0.126	0.144	42.060	0.000
12	-0.402	-0.375	77.243	0.000
13	0.193	-0.204	85.434	0.000
14	0.009	-0.139	85.452	0.000
15	0.058	0.030	86.192	0.000
16	-0.020	-0.031	86.285	0.000
17	0.060	-0.009	87.096	0.000
18	-0.102	-0.097	89.459	0.000
19	0.019	-0.016	89.537	0.000
20	0.033	-0.036	89.791	0.000
21	-0.027	-0.018	89.963	0.000
22	0.056	0.007	90.695	0.000
23	-0.022	0.062	90.804	0.000
24	-0.025	-0.242	90.952	0.000

Figure 7. Autocorrelation and partial autocorrelation plots of sequence DY12
图 7. 序列 DY12 的自相关及偏自相关图

由于序列进行了一阶十二步差分，因此 d、D 均为 1，s 为 12。通过图 7 可以初步判定 p 为 2，q 为 1。这里为了找到最优拟合模型，对 p 和 q 都分别选取 0, 1, 2 进行试验，同样，对于 P 和 Q 也分别选取 0, 1, 2 进行试验。对于模型的筛选，首先看参数显著性，如果参数不显著，则应该剔除不显著参数的自变量重新拟合模型，从低阶到高阶依次建模，最后得到 ARIMA(1,1,1) × $(0,1,1)_{12}$ 和 ARIMA(2,1,1) × $(0,1,1)_{12}$ 两个模型。进一步比较两个模型的 AIC 准则、SC 准则、HQC 准则以及拟合优度，如表 3。

Table 3. Comparison of related statistics between the two models
表 3. 两模型相关统计量的比较

数量	p	q	Q	SC 准则	HQC 准则	R^2
MOD1	1	1	1	17.3770	17.3364	0. 4598
MOD2	2	1	1	17.3740	17.3233	0. 4811

从表 3 中的数据可以看到，相较来说 MOD2 的 AIC 准则、SC 准则及 HQC 准则都更小，同时拟合优度 R^2 更大，因此认为 MOD2 更好，即选择 ARIMA(2,1,1) × $(0,1,1)_{12}$ 为最优拟合模型。

3) 模型有效性验证

要对得到的 ARIMA 模型进行有效性检验，即检查检验残差是否满足方差齐性假定，残差序列检验结果如图 8。延迟 6 阶、12 阶、18 阶、24 阶时，Q 统计量的 p 值都远远大于 0.05，因此残差序列是白噪声的，即相关信息已被模型充分提取，所建模型是有效的。

4) 模型拟合效果检验

建立了 ARIMA(2,1,1) × $(0,1,1)_{12}$ 模型后，用 2019 年的数据对其进行样本内预测，检验其拟合效果。图 9 加入了 2019 年各月新增就业人口的真实值进行对比，可以看到实际值与预测值均落在 95%的预测区间内，并且两者较为贴合。模型的预测值与预测区间如表 4。

可以看出尽管实际值与预测值有出入，但是相对误差较小，平均误差为 3.5%，预测精度较好。稳妥起见，用 ARIMA 模型对 2002 至 2019 各年度新增就业人数的月度数据进行预测，看其整体拟合效果。由于数据进行了差分，预测样本范围缩小为 2003 年 4 月至 2019 年 12 月。

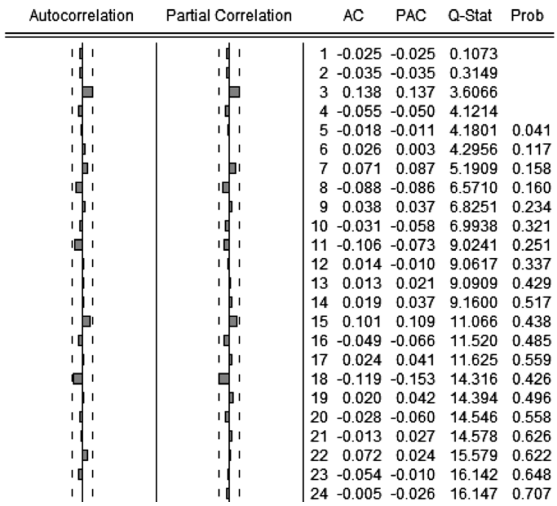


Figure 8. Autocorrelation and partial autocorrelation plots of residual series
图 8. 残差序列的自相关及偏自相关图

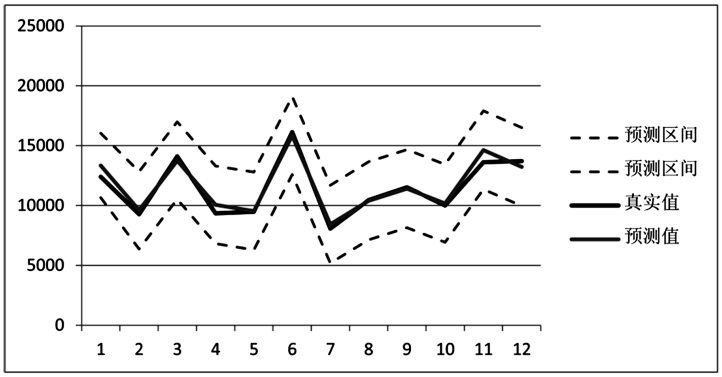


Figure 9. The deviation between the real value and the predicted value in 2019
图 9. 2019 年新增就业人口数据真实值与预测值的偏差

Table 4. System resulting data of standard experiment
表 4. 模型预测结果与真实值对比

月份	真实值	预测区间	预测值	相对误差	平均误差
1 月	12,407	[10663, 16029]	13,346	0.0757	
2 月	9281	[6376, 12838]	9605	0.0350	
3 月	14,107	[10522, 16986]	13,754	0.0250	
4 月	9342	[6814, 13305]	10,059	0.0768	
5 月	9472	[6286, 12795]	9540	0.0072	
6 月	16,127	[12588, 19110]	15,848	0.0173	0.0354
7 月	8065	[5184, 11693]	8438	0.0463	
8 月	10,459	[7137, 13661]	10,398	0.0058	
9 月	11,520	[8154, 14675]	11,414	0.0092	
10 月	9998	[6925, 13432]	10,178	0.0181	
11 月	13,616	[11346, 17911]	14,628	0.0744	
12 月	13,706	[9969, 16499]	13,233	0.0345	

经过以上步骤建立了 ARIMA(2,1,1) × $(0,1,1)_{12}$ 模型，最终计算得出的模型口径如下。

$$1-(L)\left(1-L^{12}\right)Y_t=\frac{1+0.994763L}{1-0.320162L-0.181502L^2}\left(1+0.555525L^{12}\right)\varepsilon_1$$

3.3. 预测与讨论

ARIMA 模型适用于短期的预测，因此本文对 2020 年我国的外商直接投资额进行预测。表 5 和图 10 是 2020 年贵州省城镇新增就业人数预测结果。

Table 5. Comparison of model prediction results with real values
表 5. 模型预测结果与真实值对比

月份	95%的预测区间	预测值
1 月	[10439, 15831]	13,134.92
2 月	[7044, 12761]	9902.26
3 月	[11456, 17448]	14,451.92
4 月	[6980, 13119]	10,049.87
5 月	[13430, 19609]	16,519.74
6 月	[5622, 11805]	8713.74
7 月	[7814, 14020]	10,916.59
8 月	[8829, 15045]	11,937.30
9 月	[7423, 13630]	10,526.36
10 月	[11522, 17749]	14,635.40
11 月	[10975, 17189]	14,082.11
12 月	[7130, 13214]	10,172.18

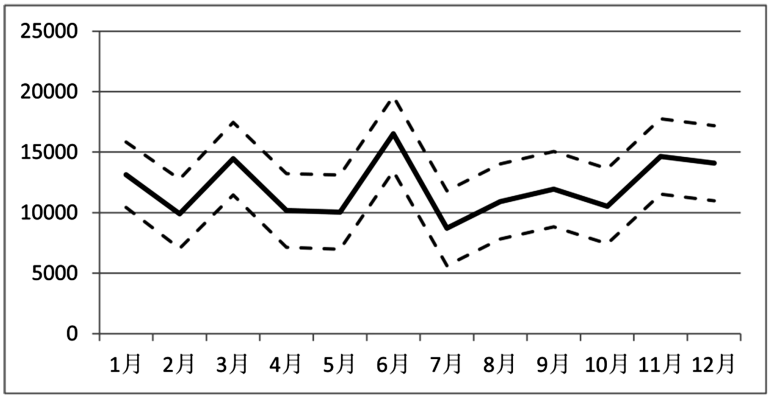


Figure 10. Curve: system result of standard experiment
图 10. 2020 年城镇新增就业人口预测结果

若不受突发事件的影响，点预测值即为新增就业人数序列的理论值。由于 2020 年初爆发了新冠肺炎，因此分析这一重要事件对就业造成的影响，能够使预测结果更贴近现实。用已知的 2020 年 1 月至 4 月贵州省实际新增就业人数，与 ARIMA 模型的预测值进行对比(见表 6 和图 11)，以此来观察并分析新冠疫情的影响程度及疫情下新增就业人口的走势。

Table 6. Comparison of model prediction results with real values
表 6. 模型预测结果与真实值对比

月份	95%的预测区间	预测值
1 月	12,680	[10439, 15831]
2 月	6740	[7044, 12761]
3 月	11780	[11456, 17448]
4 月	10140	[7130, 13214]

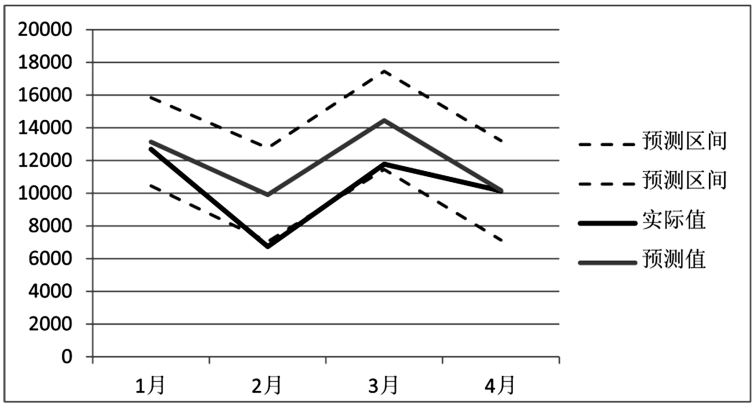


Figure 11. Comparison of the predicted and actual valuesin 2020
图 11. 2020 年新增就业人数预测值与实际值对比

1 月份，新冠肺炎尚未大范围传播，城镇就业新增人口依然维持着平稳增长的趋势，实际与预测值的相差较小，在接受范围内。2 月份，新冠疫情到了爆发期，新增就业人口开始受到极大影响；在图 11 中也可以看到 2 月的实际值超出了 95%水平下的预测区间下限。3 月份，随着中国疫情基本得到控制，与预测值相比，相对误差较上月明显减少，实际值贴近预测区间下限，回到区间范围内，可见就业状况出现一定稳定迹象。4 月份新增就业人数与预测值仅相差 0.3%，说明新冠疫情带来的影响逐渐消退，我国居民就业开始回到正轨。因此，新冠疫情这一重大随机性事件带来的影响是阶段性的、暂时的，并不会改变总体稳定的大趋势。这也说明前文 ARIMA 模型的预测结果仍是有效的。

4. 结论与展望

本文对我国 2002 年至 2019 年实际新增就业人口数据进行分析，建立 ARIMA(2,1,1) × $(0,1,1)_{12}$ 模型，样本内预测结果显示模型拟合效果较好；运用该模型对 2020 年各月份我国的外商投资额进行了预测，结果显示未来贵州省新增就业人数将保持较为稳定的季节性波动和长期增长趋势，最后进行了稳健性检验，并且对重大随机性事件造成的影响进行了分析，来进一步检验预测结果的可靠性。根据以上预测结果以及分析，建议如下：从短期看，面对疫情这类重大随机性事件，缓解“就业难”的问题是当务之急，并落实福利保障等政策以应对疫情的冲击；同时，“脱贫攻坚”工作以来，每一年度贵州省城镇居民就业人数都呈持续增长趋势，在很大程度上解决了城镇居民的基本生活问题，在一定程度上缓解了贫富差距的进一步扩大，这是贵州省“脱贫攻坚”工作的重要成果体现。

致　谢

论文最后感谢山东省自然科学基金的资助，感谢胡锋老师的指导和审稿人提出的建议。

基金项目

山东省自然科学基金(批准号：ZR2017MA012)。

参考文献

[1] 黄静. 构建居民幸福指数指标体系方法研究[D]: [硕士学位论文]. 大连: 东北财经大学, 2007: 14-25.

[2] 王春光. 贵州省脱贫攻坚及可持续发展研究[J]. 贵州民族大学学报(哲学社会科学版), 2018(3): 39-56.

[3] 刘攀, 张英. 基于统计的贵州林业扶贫攻坚成效分析[J]. 林业经济, 2017(5): 94-97.

[4] 于冰. 贵州省脱贫攻坚工作现状问题及对策[J]. 理论与当代, 2018, 411(8): 18-20.

[5] 冉茂文, 聂雪松. 贵州扶贫攻坚成效, 贫困特征及对策措施[J]. 贵州民族研究, 2000(3): 44-47.

[6] 李如是. 我国城镇就业人口变动趋势的灰色关联分析[J]. 现代商业, 2016(33): 193-194.

[7] 吴江, 王欣. 公共就业服务中存在的问题及其解决方略[J]. 城市问题, 2012(7): 59-64.

[8] 易丹辉, 王燕. 应用时间序列分析[M]. 第 5 版. 北京: 中国人民大学出版社, 2019: 154-188.

[9] 张文彤, 邝春伟. SPSS 统计分析基础教程[M]. 北京: 高等教育出版社, 2011: 89-91.