Statistics and Application 统计学与应用, 2023, 12(2), 346-354
Published Online April 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.122037

Hans汉斯

基于动态对比和聚类算法的贵州省区域CPI时序科普分析

吴明珍，晏　榛

兴义民族师范学院，数学科学学院，贵州　兴义

收稿日期：2023年3月14日；录用日期：2023年4月4日；发布日期：2023年4月19日

摘　要

区域居民消费价格指数(Consumer Price Index, CPI)是区域宏观经济分析和政策指定的重要参考指标。不同区域的CPI存在一定的差异，但目前对区域CPI的研究文献较少。文章针对贵州省及其各区域的CPI时序数据，使用对比分析与聚类的方法对贵州省各区域CPI时序的运行状态进行了分析，并普及相关科学知识。研究结果表明：贵州省各区域CPI的运行状态存在一定的差异，又从聚类结果中能看到相同的趋势。这对贵州省经济区域协调发展与物价调控政策制定提供了一定的理论依据。

关键词

贵州省区域CPI，CPI运行状态，ARIMA模型，层次聚类法

Temporal Popular Science Analysis of Regional CPI in Guizhou Province Based on Dynamic Comparison and Clustering Algorithm

Mingzhen Wu, Zhen Yan

School of Mathematical Science, Xingyi Normal University for Nationalities, Xingyi Guizhou

Received: Mar. 14th, 2023; accepted: Apr. 4th, 2023; published: Apr. 19th, 2023

Abstract

Regional Consumer price index (CPI) is an important reference index for regional macroeconomic

文章引用: 吴明珍, 晏榛. 基于动态对比和聚类算法的贵州省区域 CPI 时序科普分析[J]. 统计学与应用, 2023, 12(2): 346-354. DOI: 10.12677/sa.2023.122037

analysis and policy designation. There are some differences in CPI in different regions, but there are few studies on regional CPI at present. Aiming at the CPI time series data of Guizhou Province and its regions, this paper analyzes the running state of CPI time series of each region in Guizhou Province by using comparative analysis and clustering method, and popularizes relevant scientific knowledge. The results show that there are certain differences in the running state of CPI among different regions in Guizhou Province, and the same trend can be seen from the clustering results. This provides a theoretical basis for the coordinated development of regional economy and the formulation of price control policies in Guizhou Province.

Keywords

Guizhou Province Regional CPI, CPI Running State, ARIMA Model, Hierarchical Clustering Method

1. 引言

居民消费价格指数(Consumer Price Index, CPI)是指一篮子消费商品及服务项目的价格水平在特定时间段内的变化相对数。不仅反映通货膨胀程度，而且代表消费品及服务的总体价格对居民生活水平的实际影响。CPI 关系到诸多政策的制定与调整，是政府实施宏观调控的重要参考指标。根据地区的“宾大效应”推理，省级各地区之间、城市与农村之间的价格水平应会存在一定的差异[1]。区域 CPI 是衡量该区域总体消费价格水平变动的关键指标，也是区域宏观经济决策与核算的重要指标。

贵州省各地区经济发展程度、发展水平、资源禀赋等情况存在明显差异，要维持物价稳定，提高政策的针对性和有效性必须因地制宜的制定政策。准确分析各地区价格水平的特点和变化规律可以提高价格调控的针对性和有效性，有利于疫情下贵州省各区域经济的可持续发展，从而使贵州省经济真正实现高质量的增长。因此，研究贵州省各区域 CPI 运行状况的实证分析对政府把握地区经济动态发展趋势与差异具有重要意义。

本文结构如下：第 1 节对 CPI 相关研究内容与研究方法进行综述。第 2 节说明了研究数据的来源以及对数据进行了简单的描述性统计分析。第 3 节使用 ARIMA 模型对贵州省各区域的 CPI 运行状态数据进行了动态对比分析。第 4 节使用聚类方法进一步分析了贵州省各区域 CPI 时间序列的异同。第 5 节给出了研究结论与相关建议。

2. 文献综述

已有大量文献对 CPI 进行了相关的研究。根据研究目标或手段的不同，相关的研究可以粗略地分成两大类。第一类从动态角度出发，运用时间序列或机器学习等方法对 CPI 时序数据进行建模分析，研究 CPI 的运行状态以及 CPI 的预测。第二类是从结构视角出发，研究 CPI 波动的内因与外因以及与各影响因素之间的传导机制。在文献[2]中，王振中等提供了一个对中国 CPI 的系统分析，包括 CPI 的动态结构和可预测性，以及中美 CPI 的量化比较。

针对第一类，伊力扎提·艾热提[3]综合比较研究了中国消费者价格指数预测模型选择的问题，研究结果表明我国的 CPI 存在一定的季节性；通过条件异方差模型可以提高预测精度；通过引入外在的驱动因素来构建协整模型可以进一步保证预测的准确性。陈逸东与陆忠华[4]针对 CPI 的预测值滞后于真实值

的现象，提出一种基于卷积神经网络－长短期记忆(CNN-LSTM)深度网络的 CPI 预测模型，预测结果相较于传统方法有较小的均方根误差和平均绝对百分比误差，且预测结果的定向精度和 Pearson 相关系数显著高于传统方法。邵明振等[4]使用 BP 神经网络和 ARMA 模型对我国月度 CPI 进行了建模分析与预测，实验结果表明 BP 神经网络方法有更好的预测精度。第一类这种单变量模型忽略了其他经济变量对 CPI 预测的有用信息。

针对第二类，李博英和王璇子[5]建立了碳排放强度对 CPI 影响的向量自回归(VAR)模型，对中国 2000 年至 2020 年的时间序列数据进行了实证研究。其研究表明碳排放强度与 CPI 互为格兰杰因果，而且碳排放强度对 CPI 的影响在前期较大，随后逐渐趋于平稳。张伟[6]利用地区购买力平价(Purchase Power Parity, PPP)方法测算和推算了我国 31 个省级地区的居民消费地区 PPP，以反映各地区居民消费价格水平的差距和变动。钟妙[7]建立了 VAR 模型研究法定存款准备金率对 CPI 的影响，其结果表明法定存款准备金率在短期内干扰甚至决定 CPI 的走向，但长期无效。魏璐和钱存华[8]建立向量误差纠正模型研究了外汇储备粮、金融机构贷款额和固定资产投资额对我国 CPI 的影响。目前关于 CPI 的相关研究还包括大数据背景下的实时 CPI 指数编制方法与使用。

综上所述，目前针对 CPI 的研究方法能够有效进行预测与分析影响 CPI 运行的各种因素。但是，研究不同区域 CPI 时序的异同的文献较少。本文结合贵州省的经济发展形式，提出了研究贵州省及其各区域 CPI 时序的动态分析与聚类分析方法。

3. 数据说明

本文使用 2017 年 1 月至 2021 年 12 月(共 60 期)贵州省及其所辖九大区域的月度同比 CPI 数据(上年同月 = 100)。数据来源于贵州省统计局官网。使用 Python 绘制得到贵州省及其各区域的月度同比 CPI 时序图如图 1 所示。

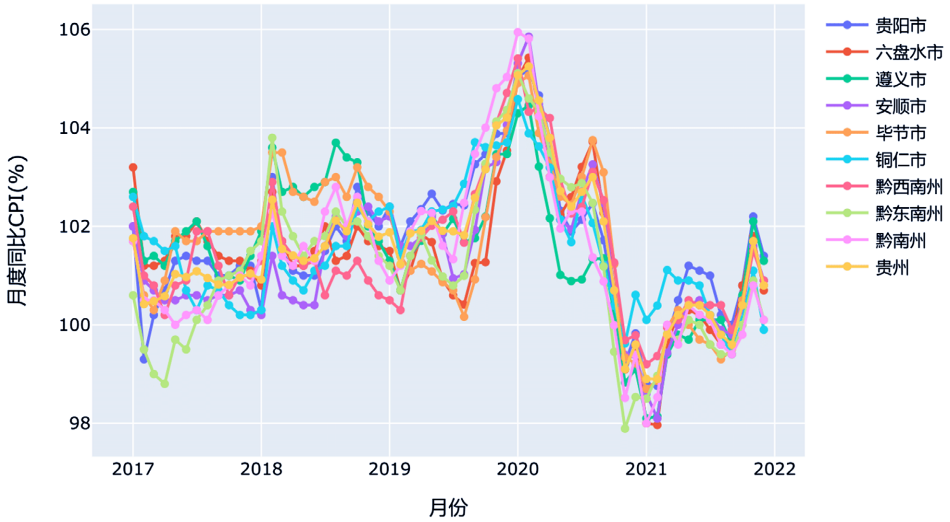

Figure 1. Time series chart of monthly year-on-year CPI in Guizhou Province and its regions
图 1. 贵州省及其各区域月度同比 CPI 时序图

4. 贵州省及其各区域 CPI 时序的动态科普分析

为了了解贵州省及其各区域 CPI 时序在动态运行上的差异，本节对各 CPI 时序进行时间序列建模分析与模型对比。

4.1. 各区域 CPI 时序的平稳性检验

从图 1 可以看出，贵州省各区域 CPI 的时序走势大致趋同，但又有区别。此外，这些时序图在直观上是非平稳的。为保证各区域 CPI 时序数据平稳性检验的可靠性，本文采用了四种单位根检验方法：ADF 检验、PP 检验、KPSS 检验与 DFGLS 检验。其中，ADF 检验、PP 检验和 DFGLS 检验的原假设均为存在单位根(即假设序列是非平稳序列)，而 KPSS 检验的原假设是不存在单位根(即假设序列是弱平稳序列)。所有假设检验均使用软件 Python 3.9 实现。假设检验结果如表 1 所示。表中所有检验的结果均是在模型设定为含有截距项(trend = 'c')且为 5%显著性水平下所得。

Table 1. The stability test results of CPI time series in Guizhou Province and other regions
表 1. 贵州省及各区域 CPI 时序的平稳性检验结果

	ADF 检验的 P 值	PP 检验的 P 值	KPSS 检验的 P 值	DFGLS 检验的 P 值
贵阳市	0.164	0.114	0.267	0.021
六盘水市	0.089	0.081	0.271	0.029
遵义市	0.443	0.188	0.053	0.137
安顺市	0.089	0.238	0.304	0.010
毕节市	0.044	0.277	0.147	0.003
铜仁市	0.394	0.260	0.220	0.133
黔西南州	0.324	0.184	0.357	0.076
黔东南州	0.213	0.343	0.277	0.035
黔南州	0.495	0.335	0.229	0.112
贵州省	0.436	0.287	0.261	0.088

虽然 KPSS 检验在所有区域的 P 值均大于 0.05，表明在该显著性水平下无法拒接原假设，即各个 CPI 时序均是弱平稳的。此外，DFGLS 检验在贵阳市、六盘水市、安顺市、毕节市和黔东南州 5 个区域的 P 值也表明相应的 CPI 时序是弱平稳的。但是，PP 检验的在所有区域序列的 P 值均拒绝了原假设，表明各个 CPI 序列是非平稳的。而且除了毕节市外，ADF 检验在其他城市的 P 值也表明相应的 CPI 时序是非平稳的。另外，通过综合对比各个区域序列的各类假设检验结果发现：六盘水市和毕节市的 CPI 序列是比较趋近于平稳的；遵义市、铜仁市和黔南州的 CPI 序列是相对比较不平稳的。

4.2. 各区域 CPI 时序的季节性检验

各区域的 CPI 时序还可能存在季节性，需要进一步对各 CPI 时序的季节性特征进行检验。如果存在季节性，则应当采用季节性模型对其进行建模预测分析。通过绘制季节性图可以直观看出各区域 CPI 时序是否存在季节性。经过实验发现，贵州省及其各区域的 CPI 时序均不存在明显的季节性特征，这从第 4.3 节使用 auto_arima 自动寻优建模的结果中得到进一步验证。贵州省的 CPI 时序的季节性图如图 2 所示。本文已省略其余 CPI 时序的季节性图。

4.3. 各区域 CPI 时序的 ARIMA 模型

根据 4.1 与 4.2 的分析结论，各区域的 CPI 时序数据适用于使用自回归积分滑动平均模型(Autoregressive Integrated Moving Average Model, ARIMA)进行建模分析，而且考虑到该模型十分简单，只需要内生变量而

不需要借助其他外生变量，是对时间序列数据进行分析和预测时比较完善和精确的算法。ARIMA 模型是 Box 与 Jenkins 于上世纪 70 年代开发的用于非平稳时间序列的建模与预测方法[9]。模型 ARIMA(p,d,q)的建模思路是：首先将非平稳的时间序列 X_t 经过 d 阶差分变换为平稳时序 Y_t，然后对 Y_t 建立自回归滑动平均模型 ARMA(p,q)：

$$Y_t = c + \sum_{i=1}^{p} \beta_i Y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{1}$$

其中 c 为常数项，$\beta_i\ (i=1,2,\cdots,p)$ 是自回归模型系数，$\theta_i\ (i=1,2,\cdots,q)$ 为滑动平均模型系数，ε 为白噪声序列，p 为自回归模型的阶数，q 为滑动平均模型的阶数。

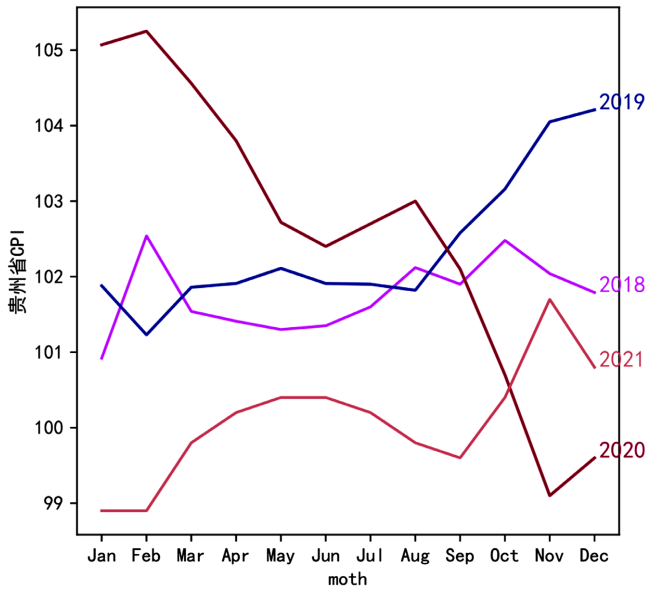

Figure 2. CPI seasonal map of Guizhou Province
图 2. 贵州省的 CPI 季节性图

由于各区域 CPI 时序在不同的平稳性假设检验理论下呈现不同的结果以及避免对所有 CPI 时序建模的繁琐过程，本文使用 Python 的第三方包 pmdarima 2.0.2 版本中的 auto_arima 函数对各区域 CPI 时序的 ARIMA 模型参数进行自动网格搜索，找出 AIC 值最低的参数。实验时均将 CPI 时序的最后 5 期作为测试数据，其余为模型的训练数据。各区域 CPI 时序的 ARIMA 模型的拟合结果如表 2 所示。

Table 2. ARIMA(p,d,q) model of CPI time series for each region in Guizhou Province
表 2. 贵州省各区域 CPI 时序的 ARIMA(p,d,q)模型

	p	d	q	AIC
贵阳市	1	0	0	135.2110
六盘水市	2	0	0	136.9980
遵义市	1	2	0	126.2090
安顺市	1	0	1	110.5350
毕节市	1	0	1	102.7620
铜仁市	5	2	0	107.9090

Continued

黔西南州	1	2	0	127.5080
黔东南州	1	0	2	117.3030
黔南州	2	2	1	129.0520
贵州省	5	3	0	133.0640

4.4. 各区域 CPI 时序的 ARIMA 模型预测

从表 2 中可知，各区域 CPI 时序的 ARIMA 模型的 AIC 值存在一定的差异，AIC 值最低的是毕节市，最高的是六盘水市。分别使用 3.3 节中拟合得到的 ARIMA 模型对后 5 期的 CPI 进行预测，并与测试集的结果进行对比计算平均绝对误差。各区域的预测结果如表 3 所示。

Table 3. Forecast results of ARIMA(p,d,q) model for CPI time series in each region of Guizhou Province
表 3. 贵州省各区域 CPI 时序的 ARIMA(p,d,q)模型预测结果

	2021-8	2021-9	2021-10	2021-11	2021-12	平均绝对误差
贵阳市	101.1118	101.2053	101.2836	101.3491	101.4040	0.6911
六盘水市	100.1583	100.5012	100.7962	101.0206	101.1828	0.5251
遵义市	100.2837	100.3830	100.5211	100.6414	100.7698	0.6069
安顺市	100.2456	100.4618	100.6367	100.7782	100.8927	0.4346
毕节市	99.8311	100.1785	100.4612	100.6913	100.8786	0.4671
铜仁市	100.3771	100.4183	100.2750	100.1783	100.0095	0.5602
黔西南州	100.3545	100.3297	100.2955	100.2656	100.2338	0.4961
黔东南州	99.6675	99.7293	100.0069	100.2303	100.4100	0.3167
黔南州	100.2456	100.4618	100.6367	100.7782	100.8927	0.6580
贵州省	100.3737	100.4946	100.4187	100.2971	100.0623	0.7255

5. 各区域 CPI 时序的聚类分析

通过第 4 节对贵州省及其各区域 CPI 时序的动态对比分析。我们发现各区域 CPI 时序在相同的建模标准下得到的最优拟合模型以及预测结果存在一定的差异。

本节使用聚类方法对各区域的 CPI 时序进一步进行分析。在研究聚类算法前，首先要明确聚类的目的。本文的聚类目的是将 CPI 时序的动态相关关系更接近的聚在一组。动态相关关系描述了时间序列的形成过程，揭示了时间序列的本质特征，可以挑选同一簇内有代表性的序列来估计整个簇内序列的特征，这在海量序列数据的研究中具有重要意义。时间序列聚类问题的定义如下：把 n 个时间序列的集合记为 $X=\left(X_1,X_2,\cdots,X_n\right)$，其中第 i 个时间序列为 $X_i=\left(X_{i1},X_{i2},\cdots,X_{in}\right)$，$T_i$ 为第 i 个时间序列的长度。聚类目标是把 n 个时间序列聚到 $C=\left\{C_1,C_2,\cdots,C_k\right\}\left(k\ll n\right)$。无需进行数据集的标准化。聚类问题的关键从技术层面角度来看侧重于对聚类算法的改进，而从理论层面来看更侧重于对距离度量的选取。

5.1. CPI 时序的 DTW 距离测度

为了确定时间序列之间的相似性及对时间序列进行分类，需要测量时间序列之间的距离。动态时间规整(Dynamic Time Warping, DTW)在数据挖掘中通常作为时间序列之间距离的测量方法。它使用动态规

划算法找到两个时间序列之间的最佳对齐。作为时间序列相似性度量，它通过允许时间序列的“弹性”变换来检测具有不同相位的相似形状，从而最大限度地减少时间偏移和失真的影响，因此获得了广泛的应用，包括语音识别、手势识别、机器人、制造和医学等。

DTW 的基本思想描述如下。任意给定两个时间序列：

$$X = x_1, x_2, \cdots, x_i, \cdots, x_{|X|}, \tag{2}$$

$$Y = y_1, y_2, \cdots, y_i, \cdots, y_{|Y|}, \tag{3}$$

它们的长度分布为$|X|$和$|Y|$。定义它们之间的一条规整路径为：

$$P = p_1, p_2, \cdots, p_k, \cdots, p_K, \tag{4}$$

其中K为该路径的长度，且满足$\max(|X|,|Y|) \le K \le |X|+|Y|$。规整路径$P$中的第$k$个元素$p_k = (p_{k1} - p_{k2})$，其中$p_{k1}$是来自时间序列的一个索引，$p_{k2}$是来自时间序列的一个索引。要求：

a) $p_1 = (1,1), p_k = (|X|,|Y|)$；

b) 若$p_k = (i,j)$且$p_{k+1} = (i',j')$，则$i \le i' \le i+1, j \le j' \le j+1$。

评价规整路径的指标被定义为规整路径的距离。规整路径P的距离被定义为：

$$Dist(P) = \sum_{k=1}^{K} Dist(p_k) = \sum_{k=1}^{K} \sqrt{(p_{k1} - p_{k2})^2}, \tag{5}$$

具有最小距离的规整路径称为最优规整路径，记为P^*。那么，时间序列X与Y之间的距离被定义为最优规整路径的距离，即：

$$DTW(X,Y) = Dist(P^*) \tag{6}$$

通过使用动态规划方法可以找到X与Y之间的最优规整路径，以获得它们之间的距离。本文使用文献[10]中提出的 FastDTW 算法求得贵州省各区域 CPI 时序的 DTW 距离矩阵如下表 4。

Table 4. DTW distance matrix of CPI time series in Guizhou Province
表 4. 贵州省各区域 CPI 时序的 DTW 距离矩阵

	贵阳市	六盘水市	遵义市	安顺市	毕节市	铜仁市	黔西南州	黔东南州	黔南州
贵阳市	0	33.6556	31.0318	26.7215	33.6140	24.6579	23.9946	31.0869	27.8917
六盘水市		0	29.5338	28.1068	29.3443	34.6049	22.0370	35.2029	30.2970
遵义市			0	33.0809	26.2316	34.2636	33.3610	39.3726	30.2969
安顺市				0	24.7544	25.9480	23.3777	28.8696	23.1760
毕节市					0	40.3020	22.5019	28.0155	25.2837
铜仁市						0	26.0256	42.1436	27.3365
黔西南州							0	32.0028	22.0140
黔东南州								0	22.9511
黔南州									0

5.2. 基于 DTW 距离的凝聚式层次聚类方法

度量两个簇间距离的方法有很多，例如最小距离、最大距离、平均距离、中心距离、类平均距离、离差平方和距离等。本文在得到表 4 中各区域 CPI 时序之间的 DTW 距离矩阵之后，使用 Ward 方法进行

迭代更新聚类矩阵。每次迭代选择距离最小的两个簇进行合并，直至最后聚成一类，并使用 Python 实现得到聚类树图，如图 3 所示。

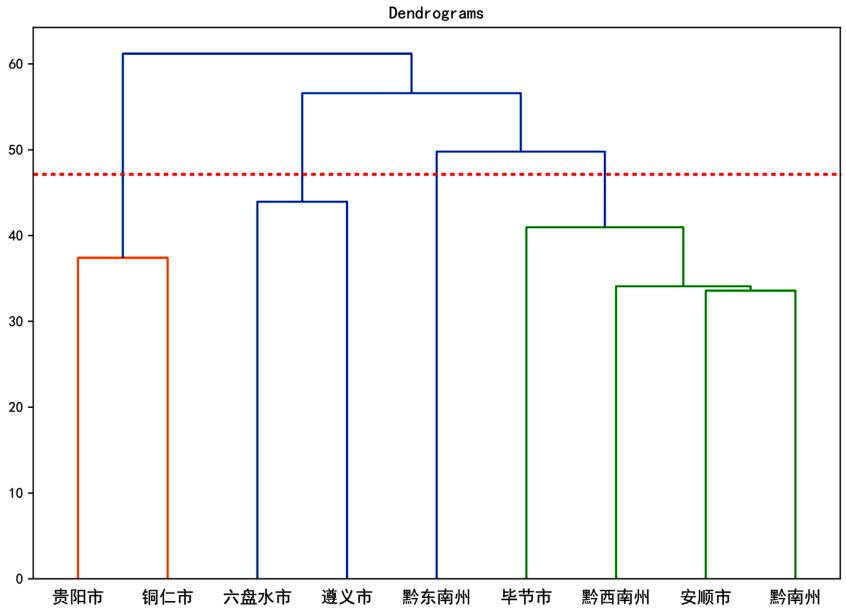

Figure 3. The cluster tree of CPI time series of each region in Guizhou Province
图 3. 贵州省各区域 CPI 时序的聚类树图

按图 3 中红色虚线划分标准，可以将贵州省各区域 CPI 时序的运行状况分为四类，分别是{贵阳市，铜仁市}、{六盘水市，遵义市}、{黔东南州}与{毕节市，黔西南州，安顺市，黔南州}。

6. 结语

本文首先使用 ARIMA 模型从动态的视角对比分析了贵州省各区域 CPI 的运行状况与变动趋势，研究发现各区域的 CPI 时序在同一建模假设下存在一定的差异。然后，基于 DTW 距离的凝聚式层次聚类方法给出了各区域 CPI 时序的聚类结果，更进一步展示了各区域 CPI 时序之间的异同。需要说明的是，本文的重点是想研究贵州省各区域 CPI 时序运行状况之间的异同，通过原始数据分析发现使用简单精确的 ARIMA 模型就能够较好的建模各区域的 CPI 时序数据，因此没有考虑使用诸如机器学习等更复杂的方法来建模，这是本文的缺陷所在，因此有待进一步探索其他建模方法。此外，研究城市 CPI 与农村 CPI 之间的差异，以及从 CPI 的影响因素的结构性差异来分析各区域 CPI 的异同也是本文的进一步研究方向。这对研究中国各地区之间 CPI 运行规律的差异有一定的启示意义。

致　谢

本文作者感谢黔西南州科技计划项目对本研究的资助，以及感谢审稿人的宝贵意见。

基金项目

黔西南州科技计划项目“贵州区域 CPI 运行状态分析研究”。

参考文献

[1] 邵明振, 陈磊, 宋雯彦. 我国月度居民消费价格指数的预测方法与应用[J]. 统计与决策, 2012(14): 30-31.

[2] 王振中, 陈松蹊, 涂云东. 中国居民消费价格指数的动态结构研究及中美量化比较[J]. 数理统计与管理, 2021(1): 109-126.

[3] 伊力扎提·艾热提. 中国消费者价格指数预测模型的选择[J]. 统计与决策, 2022, 38(4): 68-73.

[4] 陈逸东, 陆忠华. 基于卷积长短时记忆网络的 CPI 预测[J]. 计算机工程与应用, 2022, 58(9): 256-262.

[5] 李博英, 王璇子. 碳排放强度对居民消费价格指数的影响研究[J]. 统计与信息论坛, 2022, 37(10): 65-74.

[6] 张伟. 地区购买力平价与 2015-2019 年省级地区居民消费价格水平——基于 84 个主要城市的研究[J]. 统计研究, 2022, 39(10): 119-132.

[7] 钟妙. 法定存款准备金率对居民消费价格指数影响的实证[J]. 统计与决策, 2020, 36(23): 155-159.

[8] 魏璐, 钱存华. 关于对可能影响 CPI 的几个因素的研究[J]. 数理统计与管理, 2014, 33(1): 122-127.

[9] Box, G.E.P., Jenkins, G.M., Reinsel, G.C., *et al*. (2015) Time Series Analysis: Forecasting and Control. John Wiley & Sons, Hoboken.

[10] Salvador, S. and Chan, P. (2007) Toward Accurate Dynamic Time Warping in Linear Time and Space. *Intelligent Data Analysis*, **11**, 561-580. https://doi.org/10.3233/IDA-2007-11508