

基于时间序列分析的新疆红枣产量预测

苏晓燕, 桂晓蕾, 程倩, 王莹, 刘家琪

塔里木大学信息工程学院, 新疆 阿拉尔

收稿日期: 2023年10月6日; 录用日期: 2023年12月8日; 发布日期: 2023年12月15日

摘要

为了有效提高新疆红枣产量的预测精度, 文章构建了符合红枣产量的预测模型。运用GM模型与ARIMA模型分别对新疆红枣产量进行预测, 并根据二者模型特点建立了基于灰色理论与时间序列的组合预测模型对新疆红枣产量进行了预测。根据预测结果发现, 基于灰色理论与时间序列的组合模型的预测精度高, 表明了基于灰色理论与时间序列的组合模型在新疆红枣产量预测方面有较高的准确性。

关键词

时间序列, 灰色理论, 组合模型, 产量预测

Prediction of Xinjiang Jujube Yield Based on Time Series Analysis

Xiaoyan Su, Xiaolei Gui, Qian Cheng, Ying Wang, Jiaqi Liu

School of Information Engineering, Tarim University, Aral Xinjiang

Received: Oct. 6th, 2023; accepted: Dec. 8th, 2023; published: Dec. 15th, 2023

Abstract

In order to effectively improve the prediction accuracy of Xinjiang jujube yield, a prediction model suitable for jujube yield was constructed in this paper. The GM model and ARIMA model were used to predict the yield of Xinjiang jujube, and a combination prediction model based on grey theory and time series was established according to the characteristics of the two models to predict the yield of Xinjiang jujube. According to the prediction results, it was found that the combination model based on grey theory and time series had high prediction accuracy, which indicates that the combination model based on grey theory and time series has high accuracy in predicting the yield of Xinjiang jujube.

Keywords

Time Series, Grey Theory, Combination Model, Yield Prediction

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

红枣是一种温带作物，生命力顽强，能够适应强光、强辐射条件，更有耐盐碱的特质，素有“铁杆庄稼”的美称。新疆维吾尔自治区是我国红枣的主要产区之一，红枣产业已然成为脱贫致富、实现乡村振兴的支柱性产业。然而，红枣产量受多方面因素影响，构建符合红枣产量变化的预测模型，科学准确地预测红枣产量，对提升红枣产业经济效益具有重要的理论价值和实际意义。

目前在有关农产品产量的预测研究当中，单一模型以及组合模型是国内外较为广泛使用的预测方法。其中单一模型常见的方法主要有时间序列分析法、灰色预测模型、BP神经网络、VAR模型等。BP神经网络和支持向量机等预测模型存在可调参数，而可调参数的选取又是决定预测性能优劣的关键，在以往的研究中，大部分对可调参数的选取都依赖人工经验，无法实现自适应寻优。组合模型通常是将线性和非线性两种模型结合起来，通过不同的赋权方式给单一模型以权重赋值。

国内外很多学者采用ARIMA模型对数据进行预测分析，并且得到了良好的结果。王斌等[1]采用现代统计方法建立ARIMA预测模型，探讨运用时间序列模型进行订单预测是否可行。经过误差分析后得到，ARIMA(1, 1, 0)模型可以较好地反映电商平台灰枣订单的变化趋势。张艳艳等[2]以某省海域水上交通事故时间序列为例，对其季节性、周期性特征进行分析，运用ARIMA模型进行拟合预测，并与指数平滑法进行比较，以验证模型的合理性，得到了良好的效果。马晓江等[3]通过在分析河北省流行性腮腺炎流行特征的基础上，建立ARIMA模型，进行拟合度测评。此外，还有很多学者采用GM模型进行预测。任平等[4]采用灰色系统GM(1, 1)模型预测了人口、耕地以及粮食的发展趋势。李炳军等[5]基于GM(1, N)模型对我国粮食结构供需平衡进行了分析，发现我国主要粮食的需求量和产量逐年持续增长，但供需依然存在较大缺口。黄彭等[6]根据2001~2015年四川粮食产量的历史数据，建立GM(1, 1)模型，预测了2016~2018年的粮食产量。王浩等[7]以2004~2021年中国煤炭消费量为基础，建立了长期和短期2个GM(1, 1)模型，预测了3年的中国煤炭消费量。结果表明，短期的灰色模型拟合度好，精度达到99.36%。

为了提高模型预测的精度和准确度，一些学者结合各模型的优势建立组合预测模型，得到了良好的反馈。马云倩等[8]首先对LASSO-GM(1, N)、GM(1, 1)、GM(1, N)以及LASSO这四种模型的预测效果进行了比较分析，并选择采用效果良好的LASSO-GM(1, N)组合模型对2020年中国粮食产量进行了预测，预测显示，未来到2020年我国粮食产量处于稳步增长的状态。吴越等[9]分别使用Holt两参数指数平滑法和ARIMA模型预测2019~2023年长三角地区的粮食产量，并对模型进行检验，检验结果均通过，表明长三角地区的粮食产量将保持稳定增长的趋势，预测效果理想。谭满春等[10]利用ARIMA模型良好的线性拟合能力和人工神经网络强大的非线性关系映射能力，把交通流时间序列看成由线性自相关结构和非线性结构两部分组成，采用ARIMA模型对交通流序列的线性部分进行预测，用人工神经网络模型对其非线性残差部分进行预测。结果表明：组合模型的预测准确性高于各自单独使用时的准确性；组合方法发挥了2种模型各自的优势，是短期交通流预测的有效方法。

综合上述分析,不同赋权方法所产生的赋权值,在组合模型中其预测的效果也不尽相同。本文采用 GM(1, 1) 和 ARIMA 模型运用相关统计软件分析得出预测结果,结合二类模型的特点选取合适赋权方式建立组合预测模型,以提高新疆红枣产量预测精度和准度。

2. 预备知识

2.1. GM(1, 1) 预测模型

灰色系统理论是我国学者邓聚龙教授于 1982 年首先提出来的。灰色系统是指已知部分信息的样本数据所能反映的不确定性系统。灰色时间序列预测,即运用反映预测对象特征的时间序列来构造灰色模型,预测未来某一时刻的特征量。最常见的是 GM(1, 1) 预测模型,其建模方法是利用原始的数据序列作为解微分方程,对能近似满足微分模型构造条件的序列,产生具有很强的规律性的数据,并建立一种近似的差分方程模型,以此来预测未来的发展趋势。其模型建模步骤如下。

设原始非负数据序列为:

$$X^{(0)} = \{x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(n)\}$$

将原始非负序列经过一次累加生成可得一阶累加生成序列:

$$X^{(1)} = \{x^{(1)}(1), x^{(1)}(2), \dots, x^{(1)}(n)\}$$

$$x^{(1)}(k) = \sum_{i=1}^k x^{(0)}(i), k = 1, 2, \dots, n$$

作 $X^{(1)}$ 的一阶紧邻均值生成序列 $Z^{(1)}$ 得:

$$Z^{(1)} = \{z^{(1)}(2), z^{(1)}(3), \dots, z^{(1)}(n)\}$$

$$z^{(1)}(k) = \alpha x^{(1)}(k-1) + (1-\alpha)x^{(1)}(k) (k = 2, 3, \dots, n)$$

α 为权重因子,一般取 $\alpha = 0.5$, 由此可得:

$$x^{(0)}(k) + \alpha x^{(1)}(k) = b, k = 2, \dots, n$$

将 GM(1, 1) 模型的基本形式白化微分为:

$$\frac{dx^{(1)}(k)}{dt} + ax^{(1)}(k) = b$$

式中, a 是发展系数, b 是灰色作用量, a 、 b 均表示待定参数, t 表示时间。

利用最小二乘得出:

$$\hat{a} = [a, b]^T = (B^T B)^{-1} B^T Y$$

其中,

$$B = \begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \vdots & \vdots \\ -z^{(1)}(n) & 1 \end{bmatrix}, Y = \begin{bmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(n) \end{bmatrix}$$

白化微分方程式的时间响应函数为:

$$\hat{x}^{(1)}(k+1) = \left[x^{(0)}(1) - \frac{b}{a} \right] e^{-ak} + \frac{b}{a} (k=1, 2, \dots, n)$$

经过累减, 求得上式的时间响应函数为:

$$\hat{x}^{(1)}(k+1) = \left(x^{(0)}(1) - \frac{b}{a} \right) e^{-ak} + k$$

GM(1, 1)的最终模型为:

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k) = (1 - e^{-a}) \left(x^{(0)}(1) - \frac{b}{a} \right) e^{-ak}$$

2.2. ARIMA 预测模型

ARIMA 模型即差分自回归滑动平均模型, 是 ARMA 模型(自回归滑动平均模型)的一种, 属于时间序列分析。ARIMA 模型通常以 ARIMA(p, d, q)形式表示, AR 代表“自回归项”, 用 p (自回归项数)定义, 指用于预测下一个值的过去值; I 代表“差分项”, 用 d (使之成为平稳序列所做的差分阶数)定义, 指差分顺序规定的对序列执行差分操作的次数, 是指保持平稳; MA 代表“滑动平均项”, 用 q (滑动平均项数)定义, 指预测未来值时过去预测误差的数目。ARIMA 的数学形式表示为:

$$\hat{y}_t = \mu + \varphi_1 * y_{t-1} + \dots + \varphi_p * y_{t-p} + \theta_1 * e_{t-1} + \dots + \theta_q * e_{t-q}$$

其中, φ 表示 AR 的系数, θ 表示 MA 的系数。

步骤 1: 对序列绘制自相关图和偏自相关图, 进行 ADF 单位根检验, 判断是否平稳, 通过确定差分平稳确定 d ;

步骤 2: 确定 p 和 q , 拟合 ARIMA 模型。对平稳时间序列分别求得其自相关系数(ACF)和偏自相关系数(PACF), 观察绘图: ① 若 ACF 在 $q+1$ 阶突然截断, 则在 q 处截尾, 判断为 MA 序列, ② 若 PACF 在 p 处截尾, 判断为 AR 序列, ③ 若 ACF 和 PACF 皆为拖尾, 判断为 ARMA 序列;

步骤 3: 参数检验, 利用数理统计检验高阶模型新增加的参数是否近似为零, 检验模型残差的相关特性等, 利用 AIC、BIC 准则(既考虑拟合效果接近程度, 又考虑参数个数)确定模型。

2.3. 组合预测模型

组合预测模型就是将多类单项预测模型组合起来, 赋予单项模型不同权数以适当的加权平均形式得出适合的组合预测模型。基本组合预测模型为线性组合预测模型, 其基本公式为:

$$\hat{y}_t = \sum_{i=1}^m \omega_i y_{it} = \omega_1 y_{1t} + \omega_2 y_{2t} + \dots + \omega_m y_{mt} \quad (i=1, 2, \dots, m; t=1, 2, \dots, n)$$

其中, $W = (\omega_1, \omega_2, \dots, \omega_m)^T$ 为所赋的权重系数, 且 $\omega_m \geq 0, \omega_1 + \omega_2 + \dots + \omega_m = 1$, 确定权数 ω 要考虑误差平方和 SSE 最小原则: $SSE = \sum_{t=1}^n e_t^2 = \sum_{t=1}^n \left(\sum_{i=1}^m \omega_i e_{it} \right)^2 = W^T E W$, $E = (e_{it})_{m \times n} (e_{it})_{m \times n}^T$ 。

通过求解线性规划问题得出最优解 ω_0 , R_m 为元素全为 1 的 m 维行向量 E 为信息误差矩阵。

$$\begin{cases} \min SSE = W^T E W \\ s.t. R_m W = 1, W \geq 0 \end{cases} \Rightarrow W_0 = \frac{E^{-1} R_m^T}{R_m E^{-1} R_m^T}$$

3. 实证分析

选取 2011~2021 年的《新疆统计年鉴》中新疆维吾尔自治区红枣产量为研究对象, 对新疆红枣产量进行短期预测。

3.1. GM(1, 1)预测模型

在建立灰色预测模型 GM(1, 1)前, 要先对时间序列进行级比检验。若通过级比检验, 则说明该序列适合构建灰色模型, 若不通过级比检验, 则对序列进行“平移转换”, 从而使得新序列满足级比值检验。

Table 1. Grade ratio inspection
表 1. 级比检验

索引项	原始值	级比值	平移转换后序列值	平移转换后级比值
2011	105.8	-	487.8	-
2012	145.4	0.728	527.4	0.925
2013	199.87	0.727	581.87	0.906
2014	257.46	0.776	639.46	0.91
2015	305.43	0.843	687.43	0.93
2016	288.85	1.057	670.85	1.025
2017	315.9	0.914	697.9	0.961
2018	361.19	0.875	743.19	0.939
2019	372.76	0.969	754.76	0.985
2020	381.25	0.978	763.25	0.989
2021	345.41	1.104	727.41	1.049

表 1 展示了序列值和级比值, 平移转换后序列的所有级比值都位于区间 $\left(e^{-\frac{2}{n+1}}, e^{\frac{2}{n+1}}\right)$ 即(0.846, 1.181)内, 说明平移转换后序列适合构建灰色预测模型。

Table 2. Model parameters
表 2. 模型参数

发展系数 a	灰色作用量 b	后验差比 C 值
-0.033	557.219	0.122

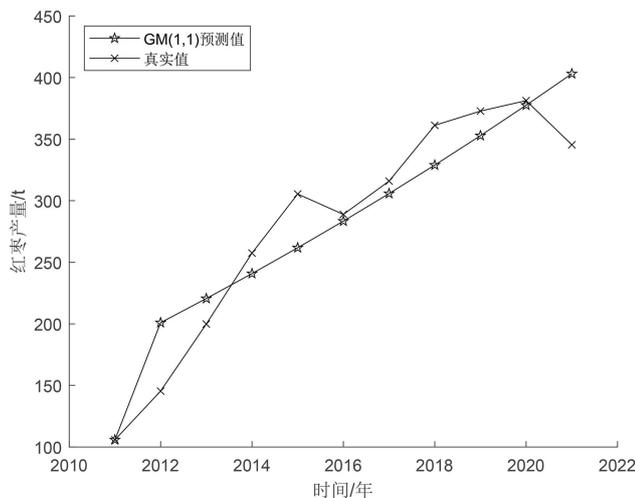


Figure 1. GM(1, 1) model prediction fitting effect diagram
图 1. GM(1, 1)模型预测拟合效果图

灰色预测模型要经过检验才能判定其是否合理, 只有通过检验的模型才能用来做预测, 系统主要通过后验差比 C 值来对灰色预测模型进行检验。从表 2 中看出, 后验差比值 C 值 0.122 小于 0.35, 一般后验差比值 C 值小于 0.35 则模型精度高, C 值小于 0.5 说明模型精度合格。灰色预测模型的预测值的平均相对误差为 9.77%, 一般情况下小于 20% 即说明拟合良好。GM(1, 1) 模型预测拟合效果如图 1 所示。

3.2. ARIMA 预测模型

根据 ARIMA (差分自回归滑动平均模型) 的具体步骤要求, 要先对数据进行 ADF 检验, 如表 3 所示, 分析的数据为不平稳数据。

Table 3. ADF inspection table

表 3. ADF 检验表

变量	差分阶数	t	P	AIC	临界值		
					1%	5%	10%
yield (t)	0	-2.534	0.107	66.499	-4.939	-3.478	-2.844
	1	-1.176	0.684	60.32	-4.473	-3.29	-2.772
	2	-3.063	0.029	63.819	-4.665	-3.367	-2.803

由表 3 可知, 原序列 P 值为 0.107 大于 0.05, 水平上呈非显著性, 接受原假设, 原序列为非平稳时间序列。差分为一阶时, 显著性 P 值为 0.684 大于 0.05, 水平上呈现非显著性, 接受原假设, 该序列为非平稳时间序列。在差分为二阶时, 显著性 P 值为 0.029 小于 0.05, 水平上呈现显著性, 拒绝原假设, 该序列为平稳时间序列。将不平稳数据进行差分运算, 得出自相关图(图 2)与偏自相关图(图 3)。在差分阶数为二阶时, 根据数据显示结果得出: 显著性 P 值为 0.029 小于 5%, 水平上呈现显著性, 拒绝原假设, 该序列为平稳的时间序列。

由图 2 可知, 一阶自相关系数明显大于 2 倍标准差范围, 且自相关系数衰减为小值波动的过程非常迅速, 可以判断自相关图为拖尾。第 2 个至第 10 个间隔下的序列自相关系数都在置信区间内, 说明序列是不相关的, 但第一个间隔的相关系数超过置信区间, 说明序列不是完全不相关的。由图 3 可看出, 第一个偏自相关系数明显大于 2 倍标准差范围, 之后的偏自相关系数都落在 2 倍标准差范围以内, 且由非零偏自相关系数衰减为在零附近小值波动的过程非常迅速, 我们可判断出偏自相关图截尾。由图 2 和图 3 与模型选择原理可以得出 ARIMA 模型的 p 、 d 、 q 值分别为 2、0、0。



Figure 2. Second order auto-correlation graph

图 2. 二阶自相关图



Figure 3. Second order partial auto-correlation graph
图 3. 二阶偏自相关图

Table 4. ARIMA(2, 0, 0) model inspection table
表 4. ARIMA(2, 0, 0)模型检验表

ARIMA 模型(2, 0, 0)检验表		
项	符号	值
	Df Residuals	8
样本数量	N	11
Q 统计量	Q6 (P 值)	0.394 (0.530)
信息准则	AIC	115.287
	BIC	116.879
拟合优度	R^2	0.708

由表 4 可知，最优模型 ARIMA(2, 0, 0)得出本次模型检验结果，包括样本数、自由度、Q 统计量和信息准则、模型的拟合优度。从 Q 统计量结果分析可以得到：Q6 在水平上不呈现显著性，不能拒绝模型的残差为白噪声序列的假设，同时模型的拟合优度 R^2 为 0.708，模型表现较为良好。ARIMA(2, 0, 0)模型预测值拟合效果如图 4 所示。

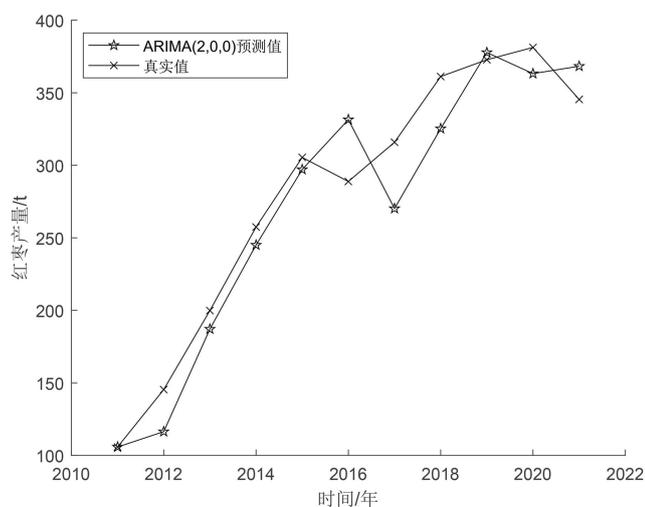


Figure 4. ARIMA(2, 0, 0) model prediction value fitting effect diagram
图 4. ARIMA(2, 0, 0)模型预测值拟合效果图

由图 4 可知,模型被拟合后进行了残差分析,通过绘制模型的残差(观测值与模型预测值之间的差异)图(如图 4 所示),可以清晰地看到自相关函数图显示残差在滞后阶数上显著不为零,说明残差之间存在自相关。残差分析是时间序列模型调优的重要步骤,有助于确保模型的可靠性和准确性。

3.3. 组合预测模型

选取 2011~2021 年新疆红枣产量数据作为研究对象,采用单一模型与组合模型分别预测其产量(见表 5)。其中,GM(1,1)模型的最小相对误差为 0.977%,最大相对误差为 38.198%,平均相对误差为 9.77%;ARIMA 模型的最小相对误差为 1.34%,最大相对误差为 19.94%,平均相对误差为 7.81%;组合预测模型的最小相对误差为 2.5%,最大相对误差为 14.5%,平均相对误差为 6.52%。比较这三者的平均相对误差,其中组合模型的平均相对误差值最小,GM(1,1)模型的平均相对误差值最大。因此,可知组合预测模型拟合精度明显高于单一预测模型,极大地改进了单一预测模型,提高了预测精度。

Table 5. Predicted values and their predictive effects of each model

表 5. 各模型预测值及其预测效果

年份	实际产量(t)	灰色预测模型		ARIMA 模型		组合预测模型	
		预测值	相对误差	预测值	相对误差	预测值	相对误差
2011	105.8	105.8	0	105.8	0	105.8	0
2012	145.4	200.941	38.198	116.4	19.94	166.48	14.5
2013	199.87	220.544	10.344	187.1	6.39	206.91	3.52
2014	257.46	240.807	6.468	245.03	4.83	242.53	5.8
2015	305.43	261.751	14.301	297.1	2.73	276.16	9.58
2016	288.85	283.4	1.887	331.5	14.77	303.01	4.9
2017	315.9	305.777	3.205	270.2	14.47	291.28	7.79
2018	361.19	328.906	8.938	325.3	9.94	327.44	9.34
2019	372.76	352.813	5.351	377.76	1.34	362.98	2.62
2020	381.25	377.524	0.977	363.2	4.73	371.69	2.5
2021	345.41	403.065	16.692	368.3	6.63	388.89	11.18
2022		429.466		305.1		394.37	
2023		456.755		267.1		402.25	
2024		484.961		235.6		437.91	

通过构建 ARIMA(差分自回归滑动平均模型)和 GM(1,1)灰色预测组合预测模型,预测出 2022 年到 2024 年的产量,组合模型预测值拟合效果如图 5 所示。由图 5 可知,组合模型预测值拟合图以一个直观的方式来展示组合模型的整体性能,实际观测值和组合模型的预测值绘制在了同一张散点图上,横轴表示实际观测值,纵轴表示组合模型的预测值,每个点代表一个数据点,观察散点分布与拟合线的关系。如图 5 所示,散点集中在拟合线附近并且呈现出近似 45 度角的趋势,说明组合模型的预测与实际观测值拟合较好。

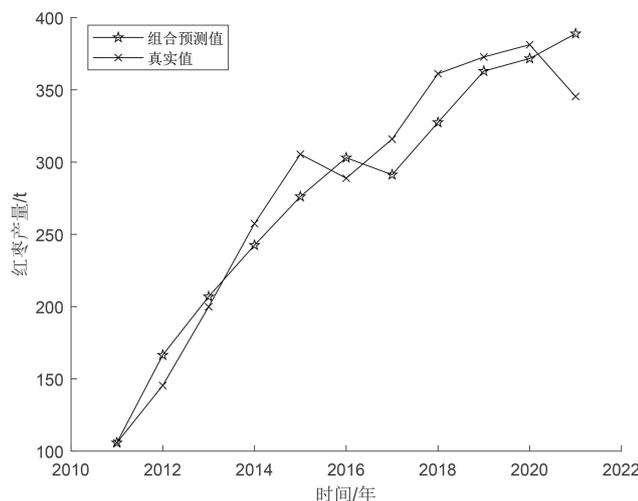


Figure 5. Combined model prediction value fitting effect diagram
图 5. 组合模型预测值拟合效果图

4. 结论

本文选取了 2011~2021 年的《新疆统计年鉴》中新疆维吾尔自治区红枣产量年度数据为研究对象，对新疆红枣产量进行了短期预测。建立了单预测模型 GM(1, 1)模型与 ARIMA 模型，分别对新疆红枣产量进行预测，单独采用 GM(1, 1)模型与 ARIMA 模型进行预测时，由于各个单一预测模型存在自身条件的局限性，不能全面准确地契合原数据本身，从而降低了预测结果的精确度。建立结合 GM(1, 1)模型与 ARIMA 模型的组合预测模型，经过实证分析结果得出组合预测模型的精度和准确度远高于单一预测模型，模型的预测拟合效果好。因此，这种组合预测模型用于新疆红枣产量预测是可行的，新疆维吾尔自治区相关部门应该根据预测相应结果积极采取一定的调控措施，科学、良好、稳步地推动新疆红枣产业迈上新台阶。

基金项目

塔里木大学大学生创新创业训练项目(编号：2023172)。

参考文献

- [1] 王斌, 杨抒, 贾清, 等. ARIMA 模型在电商平台新疆灰枣订单预测中的应用研究[J]. 福建电脑, 2019, 35(11): 5-8.
- [2] 张艳艳, 刘晓佳, 熊子龙, 等. 基于 ARIMA 模型的水上交通事故预测[J]. 中国水运(下半月), 2017, 17(2): 51-54.
- [3] 马晓江, 周然, 邹显东, 高贵军. 基于 ARIMA 模型探讨 2008-2021 年河北省流行性腮腺炎流行趋势[J]. 预防医学情报杂志, 2023, 39(11): 1299-1305.
- [4] 任平, 周介铭, 王广杰, 等. 四川省人口耕地和粮食可持续发展研究[J]. 国土与自然资源研究, 2005(4): 30-31.
- [5] 李炳军, 杨卫明. 基于灰色区间预测和 GM(1, N)模型的我国粮食供需结构平衡分析[J]. 江苏农业科学, 2019, 47(18): 325-329.
- [6] 黄彭, 郝妙, 杜永华, 等. 基于 GM(1, 1)模型的四川粮食产量预测研究[J]. 农学学报, 2017, 7(10): 96-100.
- [7] 王浩, 成玮. 基于 GM(1, 1)模型煤炭消费量预测研究[J]. 能源与节能, 2023(5): 51-53+57.
- [8] 马云倩, 郭燕枝, 等. 基于 LASSO 与 GM(1, N)模型的中国粮食产量预测[J]. 干旱区资源与环境, 2018, 32(7): 30-35.
- [9] 吴越, 张焕明. 基于 Holt 两参数指数平滑法和 ARIMA 模型的长三角粮食产量的预测[J]. 武汉轻工大学学报, 2020, 39(1): 30-36.
- [10] 谭满春, 冯萃斌, 徐建闽. 基于 ARIMA 与神经网络组合模型的交通流预测[J]. 中国公路学报, 2007, 20(4): 118-121.