

**项目学术报告**

基于我国粮食经济发展战略背景下的小麦价格ARIMA-LSTM-XGBoost组合预测模型研究

**学生姓名：苏晓钰**

**指导老师：林少萍**

**职称：讲师**

**暨南大学**

**2023年3月**

# 摘 要

国家粮食价格既是调节粮食产品供求关系的信号和手段，也关系到老百姓生活成本和农民收入水平。中国的小麦作为三大主粮之一，年产量超1亿吨，占据世界市场近25%，处于世界粮仓的重要战略地位。然而自2018年起，中美贸易摩擦不断，为中国小麦进口带来更多不确定性；2020年新冠疫情袭来，国内农产品生产受阻，小麦价格一度上涨；而国际疫情反复不断引起的市场波动、俄乌冲突动摇了乌克兰作为 “欧洲粮仓”的重要地位，推动国际小麦价格，加剧了国际市场价格波动的风险。而国内研究的研究范围和研究方法存在局限性，大多仅使用传统的时间序列分析进行预测，且很少将国际动荡事件纳入影响因素，结论过时。在当前国家粮食经济发展战略背景下，构建更精准的小麦期货价格预测模型有利于更有效地反应小麦市场，保障小麦最低收购价政策顺利实施。

本项目基于《中国农产品价格调查年鉴》、《中国统计年鉴》与国家粮油信息中心的公开数据，从小麦国际和国内市场角度出发，探求影响小麦价格的重要因素，旨在对小麦发展趋势做出准确的预测。首先根据训练集数据，基于随机森林特征重要性排序构建了全面的小麦期货价格的影响因素指标体系，并构建出传统的时间序列分析模型与单变量的的LSTM模型与单变量XGBoost模型。为了进一步构建更准确的小麦期货价格预测模型，我们基于所建立的影响因素指标体系，构建多变量的LSTM模型与多变量XGBoost模型。最后，基于Stacking集成学习算法改进现有的预测模型，考虑构建ARIMA、多变量的XGBoost模型和多变量的LSTM模型的组合模型：将三组模型的预测值作为支持向量机回归SVR模型的输入，并采用贝叶斯优化算法对SVR模型进行调参，解决影响因素的未来值的获取问题并显著提高预测精度。综合组合模型与单一模型、单变量模型与多变量模型的预测结果，从绝对误差、相对误差、平均相对误差等指标比较模型，并选择出预测效果最好的模型。基于该模型的有效预测结果，提出相应政策建议，有助于及时调节市场供求关系，稳定粮食经济；将该模型进一步推广到农产品价格预测领域，对我国农业经济发展有重大意义。

**关键词：** 时间序列预测,ARIMA,LSTM,XGBoost,随机森林

# 1 选题背景

## 1.1国家战略

中国是粮食生产大国，也是消费大国，粮食问题是关系全局的重大战略问题，而粮食价格是最主要的关注点之一。党的十九大报告中强调：要确保国家粮食安全。中央一号文件连续十二年关注“三农问题”，并提出“发展农产品期货交易”。而小麦作为三大主粮之一，在中国年产量和消费量都居世界前列。2006年起，我国对小麦实行最低收购价政策，稳定主粮价格、改善供求关系以保障农民利益。在近十年的国家政策扶持下，国产小麦产量提升、粮食库存充足，有效保障市场需求，小麦现货价格稳步上涨。

## 1.2 国内外市场

中国小麦年总产量超1亿吨，自给率达到95%以上，在国民经济中占比巨大。2021年小麦总产量达13694.6万吨，增长率2%。随着人口增长和经济发展，小麦需求量逐年上涨。得益于临储和“托市”政策，我国小麦价格从2009年1月的1790元/吨上涨至2013年9月的2519.29元/吨。2020至今，粮食稳定成为国家安全战略部署重要手段，国内小麦减产、储备性需求激增导致价格明显上涨。国内中筋小麦品种占比超55%，是国产小麦重要来源。进口成为补充我国优质小麦缺口的重要途径。

小麦是国际农产品贸易中最为活跃、用途最广的农产品之一，其种植分布、消费及总贸易额占世界粮食作物第一。自2001年起，我国逐步从小麦净出口国转变为净进口国，主要进口优质小麦和饲料用小麦以调剂粮食结构。2021年小麦进口量977万吨创新高。长期以来，美国、加拿大和澳大利亚进口小麦占我国进口量八成以上，依存度高。随着一带一路倡议提出，中国和俄罗斯、乌克兰、哈萨克斯坦的小麦贸易关系逐步建立，寻求多元的进口途径，降低对少部分国家的粮食依存度。

## 1.3 其他影响

2020新冠肺炎疫情袭来，粮食安全成为关注焦点。疫情爆发初期，国内受疫情影响，交通封锁、停工停产等使粮食运输受限导致原料短缺、面粉生产加工停摆。随着疫情转移到国外，2020年7 月以后，国际谷物价格进入了上升通道。受疫情影响，各国封锁下进出口贸易受阻，外部供给环境恶化；谷物出口大国担忧粮食供应，相继限制粮食出口。由于国内小麦现价与国际价格紧密关联，国际疫情引起的市场波动往往对我国小麦生产销售市场产生影响。

# 2 方案论证

## 2.1 项目简介

本项目基于《中国农产品价格调查年鉴》中2009年1月4日至2021年12月17日共3255个样本我国小麦期货价格序列的日数据，以及《中国统计年鉴》、国家粮油信息中心关于农业经济投入、人力投入、我国财政收支、小麦对外贸易、小麦国际进出口量，中国人民银行、国际清算银行、国际货币基金组织关于国家经济政策、国内外经济环境、现货基本条件等多个影响因素，建立数学模型预测小麦期货价格。

## 2.2 理论依据

### 3.1 ARIMA模型

整合自回归滑动平均模型ARIMA(Autoregressive Integrated Moving Average)是自回归和滑动平均模型的综合，其中心思想是运用数学模型近似描述一个随机的时间序列，可以处理非平稳性的、具有线性趋势和季节、循环成分的时间序列。

建立一个ARIMA模型需要解决以下3个问题：

（1）将非平稳序列转化为平稳序列。

（2）确定模型的形式。即模型属于AR、MA、ARMA中的哪一种。

（3）确定变量的滞后阶数。

ARIMA模型的建模预测主要包括以下步骤：

1. 通过时序图、ACF和PACF图结合来判断序列是否平稳，同时通过ADF单位根检验法构造ADF统计量，进行平稳性检验。
2. 当序列不平稳时，为例消除显著的线性趋势和季节趋势的影响，我们对时间序列进行平稳化处理，如差分法、数据的分解与平滑、对数化等。
3. 通过自相关系数图(Autocorrelation Function, ACF)和偏自相关系数图(Partial Autocorrelation Function, PACF)定阶法，观察两图的截尾和拖尾情况，初步确定p(P)和q(Q)的阶数，并结合AIC信息准则找到适合模型的最佳阶数P，Q的值。AIC准则是由K-L信息量推导而来，其一般形式为：



L是极大似然函数，m是待估参数个数，AIC值越小代表模型效果越好

表3-2 ARIMA模型定阶原则

|  |  |  |  |
| --- | --- | --- | --- |
| 模型定阶 | AR(p) | MR(q) | ARMA(p,q) |
| 自相关系数() | 拖尾 | P阶截尾 | 拖尾 |
| 偏自相关系数() | Q阶截尾 | 拖尾 | 拖尾 |

1. 考虑到小麦期货价格序列的季节效应、长期趋势效应和随机波动之间存在复杂的交互影响关系，简单的ARIMA模型不足以提取其中的相关关系，因此我们采用乘积季节模型对模型进行优化，根据最小AIC与似然比统计量确定模型。
2. 通过对残差序列的随机性检验与正态性检验评价模型是否拟合恰当，本项目将采取观察残差标准差图、残差序列Q-Q图、Ljung-Box法来检验模型的显著性。Ljung-Box法白噪声检验的假设条件为：

原假设

备择假设:至少存在某个

检验统计量QLB取为：

综上所述，本研究建立的ARIMA模型的应用步骤主要包括平稳性判别、模型定阶与识别、参数估计、模型的优化选择、模型的显著性检验、模型预测六个步骤。

****

图3-1 ARIMA模型构建流程图

ARIMA模型能较准确地预测小麦现货价格的时间序列整体趋势，但对局部拟合效果不同，且部分拟合精度较低。该模型在外推时不方便加入外生变量，未考虑结构变动带来的影响，因此仅用ARIMA模型可能导致未来的预测效果欠佳。

### 3.2 LSTM神经网络

引入LSTM模型可以重新设置RNN的隐藏层模块结构，实现了对信息的长期记忆。LSTM网络主要是通过引入门结构：遗忘门（forget gates）、输入门（input gates）和输出门（output gates），对信息进行保护和控制。

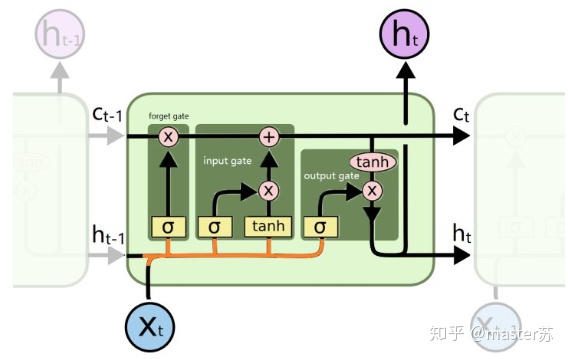


图3-3 LSTM网络单元结构

遗忘门是对于是否保留或丢弃上一时刻的、到当前时刻的状态做出抉择。通过接收上一个状态的输出值和当前时刻的输入，使用sigmoid激活函数得到0-1之间的值，其中1表示全保留，0表示全忘记，具体计算公式为：



输入门的作用是判断要留下多少当前的输入信息，共同更新当前时刻的单元状态。首先第一层由sigmoid层判断需要更新的输入，即



第二层由tanh层创建可以更新到当前状态的候选信息，即



根据遗忘门和输出门两个门结构的计算可以得到记忆细胞的更新值：



输出门的功能是决定当前时刻中可以输出到的部分，由通过sigmoid层得到的中间变量与经过tanh层处理后得到决定：





对得到的训练集预测值进行反归一化可以得到2009.07-2021.07年真正的预测数据，并绘制相应的拟合曲线。

### 3.3 XGBoost模型

极端梯度提升XGBoost(Extreme Gradient Boosting)模型是一种分布式高效梯度提升算法，可解决分类、回归等机器学习问题。它是一种集成模型，由多棵CART决策树作为子模型构成，采用Gradient Tree Boosting算法进行集成学习，最终结果为全部决策树的预测值累加之和。

### 3.4 Stacking集成学习

集成学习(Ensemble learning)是通过整合数个有差异的个体学习器学习得到的结果，从而获得一个优于单个个体学习器预测性能的集成模型。集成学习模型可以分为同质集成和异质集成。Stacking集成学习方法属于异质集成模型，它对不同的学习模型进行组合，并采用了元学习模型代替了投票法。在对个体学习器进行选择时，应尽量选择预测效果好且异质性较强的模型，这样通过元学习模型的集成学习才容易取得更好的预测成果。

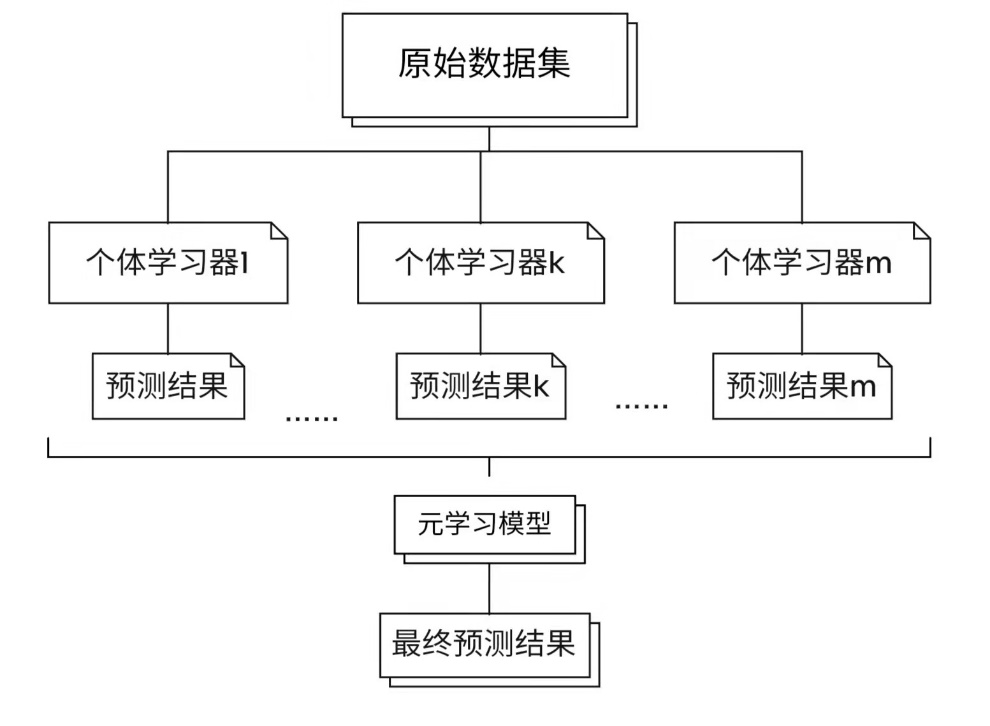


图3-5 集成学习流程图

## 3.5 贝叶斯优化算法

在模型拟合时，选择合适的参数以实现高精度并非总是容易的，既可能出现拟合过度，也可能出现拟合不足，不同的参数设置可能会导致模型性能上的显著差异。贝叶斯优化算法正是为了寻找最优的参数组合而提出的，近年来，它在机器学习超参数调优方面得到了广泛的应用。

与网格搜索和随机搜索优化算法相比，贝叶斯优化算法能找到最优的超参数集，具有更高的搜索效率山，对非凸问题进行调参的结果仍然是稳健的；同时，它能在开采和勘探之间找到一个平衡点来避免算法陷入局部最优，使得模型的性能达到全局最优21。

基于python软件的hyperopt模块实现贝叶斯优化算法主要包含以下四个部分：

(1)定义目标函数。目标函数即为我们想要使其达到最小化的对象，一般使用当前参数下的数据集的损失函数。

(2)设定参数搜索空间。在目标函数确定后，应对参数的搜索范围和分布进行设定，使算法在该区间下寻找是定义的目标函数达到最小的参数最优组合。

(3)优化算法的选择。优化算法的作用为在构造替代函数后选择下一组超参数组合，基于目标函数进行评估，主要包括：随机搜索，模拟退火以及TPE算法等。

(4)结果的历史数据。在此部分中，能够记录并存储历史选择的超参数组合和所定义的目标函数值，也就是我们想要最小化的函数值。

## 2.3 研究意义

**（1）调节市场供求关系**

价格受供求关系影响，又反作用于供求关系。过去20年，我国小麦产量稳步提升，基本满足国民消费需求，粮食库存充足，同时受政策影响价格涨幅较大。由于国内小麦生产从成本高，国家通过小麦最低收购价补贴、临储麦政策“托市”，维持了小麦供应量，从而保障消费需求。

在此期间，我国粮食期货市场日臻成熟，小麦期货价格成为市场价格“风向标”。作为大宗商品，小麦期货为农民和供应链下游的加工商、贸易商提供套期保值的功能。通过期货市场和现货市场的信息传递，有效反应市场当前和预期供求关系。合理的价格预测，一方面，指导农民调整小麦种植结构、合理安排售粮时间，保障粮食供给、控制供求不平衡带来的损失；另一方面，对于供应链下游，有助于生产商、销售商根据预期价格趋势进行采购、生产和存货管理。[2] 通过有效价格预测，减少卖方与市场的信息不对称，提高整条小麦供应链效率、保障生产者利益，满足市场需求，构建健康的主粮市场。

**（2）规范国内小麦定价**

2006年实行小麦最低收购价以来，国产小麦价格增长近一倍。由于长期以来粗放型的生产方式导致我国小麦种植成本高于国际水平，为保障国家粮食安全，国家政策补贴在一定程度上扭曲小麦价格。2016年对于粮价补贴政策的改革在一定程度上提升了市场对于小麦价格的调节能力，此后我国小麦价格一直在小范围波动上涨，更加有效地反应市场供求关系。

此外，作为粮食进口大国，我国小麦贸易逆差大、买方市场势力相对较弱，对于粮食定价影响力低。[14] 相比于出口大国对于小麦种植的扶持力度，我国还有一定距离。尽管国内外价差有缩小趋势，但今年三月以前，进口小麦价格一直优于国内，国产小麦价格及质量竞争力较低。对小麦价格的有效预测，是国家加快推进小麦和稻谷定价机制、进行补贴政策和收储体制改革的重要依据；与此同时，帮助人们有效应对国际粮价波动带来的风险，促进国产小麦价格与国际化、标准化接轨，提升我国小麦产业国际影响力。

**（3）保障国家粮食安全**

后疫情时代，在小麦价格波动剧烈、市场风险加剧的整体经济大环境下，准确预测小麦现货价格有利于降低外部环境对于粮食供应、粮食价格的冲击，优化粮食储备。尤其是当下疫情和政治等不确定性因素加剧，小麦现货市场动态是外界环境的“晴雨表”。

作为金融期货市场的物质基础，小麦期货市场围绕现货价格波动。近期小麦期货价格的暴涨，在一定程度上反映了主粮，这类大宗商品，其避险属性突出；此外，众多投机者借助国际事件涌入小麦期货市场以期进行无风险套利，同时借机推高国际利率，收割高额回报。[15] 这一特殊时期，小麦价格的合理预测有助于稳定现货价格，缓解全球性恐慌造成的价格虚高，平衡期现货市场从而避免投机行为的出现。小麦作为国家主粮，其作用应该是满足人们口粮需求、确保社会生产有序，而非投机者哄抬物价获利的工具。因此，掌握未来价格趋势的对于小麦发挥粮食作用、国家粮食供应稳定具有重要意义。

# 3 研究方法

主要为文献研究、实证分析等方法。

第一部分：构成小麦期货价格的影响因素指标体系。

计算Pearson相关系数度量变量之间先行相关关系，变量之间的非线性相关程度通过随机森林特征重要性排序，最终筛选出少数具有影响力的因素，构成预测小麦期货价格的指标体系。

第二部分：基于所构建的小麦期货价格的预测指标体系，建立模型进行预测。

（1）通过传统的ARIMA模型建立经典的单变量时间序列模型；

（2）建立预测市场现货价格的单变量BP神经网络模型；为了提升模型预测效果，根据由随机森林提取的特征构建影响因素指标体系，构建多变量BP神经网络模型；

（3）将小麦价格序列分解为线性趋势和非线性趋势，构建单变量的ARIMA-BP组合预测模型。

第三部分：评价单一预测模型与组合模型的预测效果，并进行最优选择。

通过计算得到测试集的误差、相对误差以及平均相对误差，并比较单一模型（ARIMA、单变量BP神经网络模型），组合模型（ARIMA-BP组合预测模型），多变量模型（多变量BP神经网络模型）等预测模型下的指标，最终选出最适合的预测模型，并应用于我国小麦市场发展趋势预测。

## 3.5技术路线图

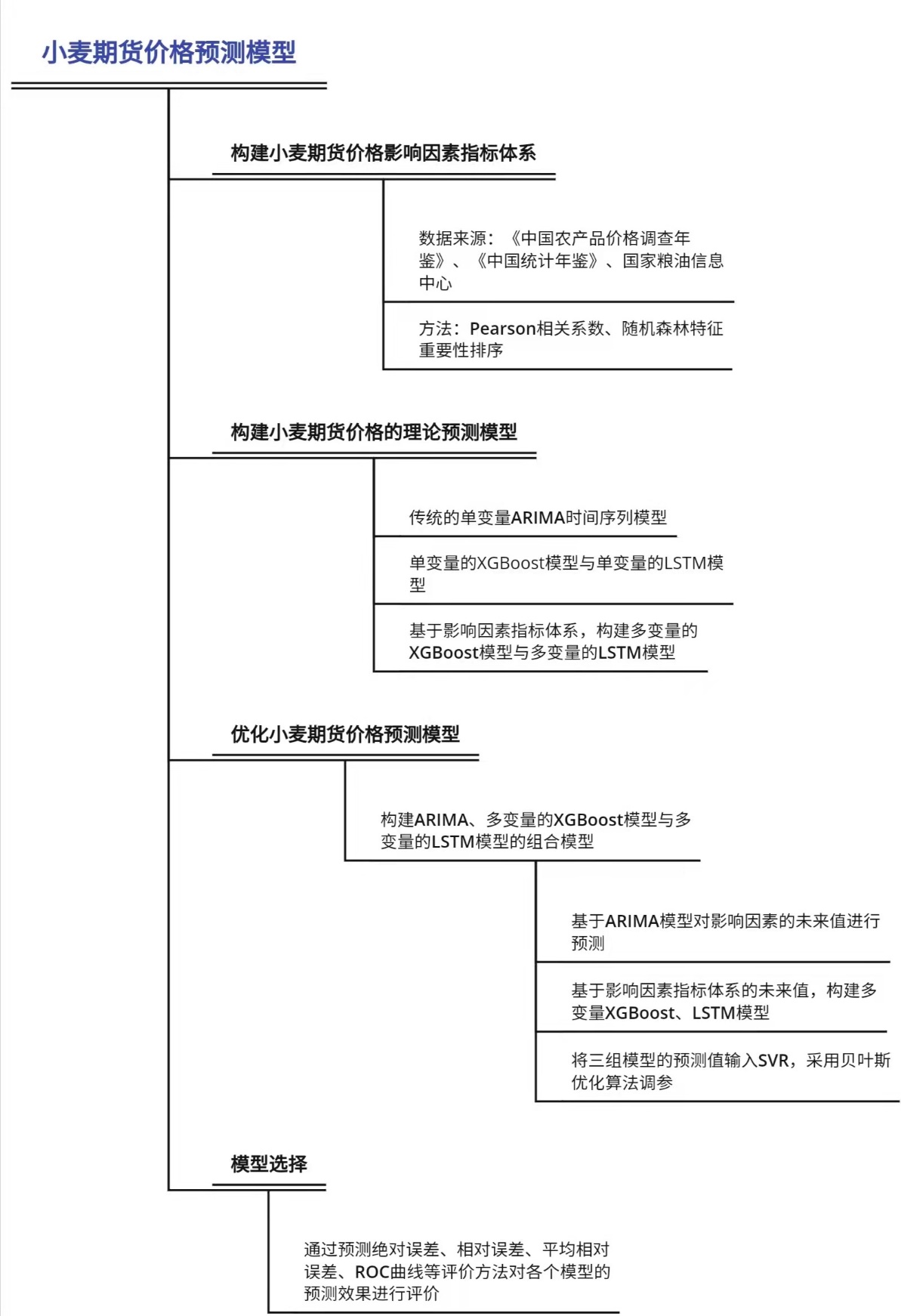
****

图3-5 小麦期货价格预测模型技术路线图

# 4 研究结果

# ARIMA时间序列模型小麦价格预测

本次研究将2009年1月至2021年4月作为训练样本（共148个月样本），2021年5月至2021年9月作为测试样本（共5个月样本）。

## 平稳性检验

### 时间序列平稳性判别

平稳序列有着均值、方差为常数，不受其他因素的影响而变化的特点。本文首先将通过绘制时间序列图观察序列的平稳性。

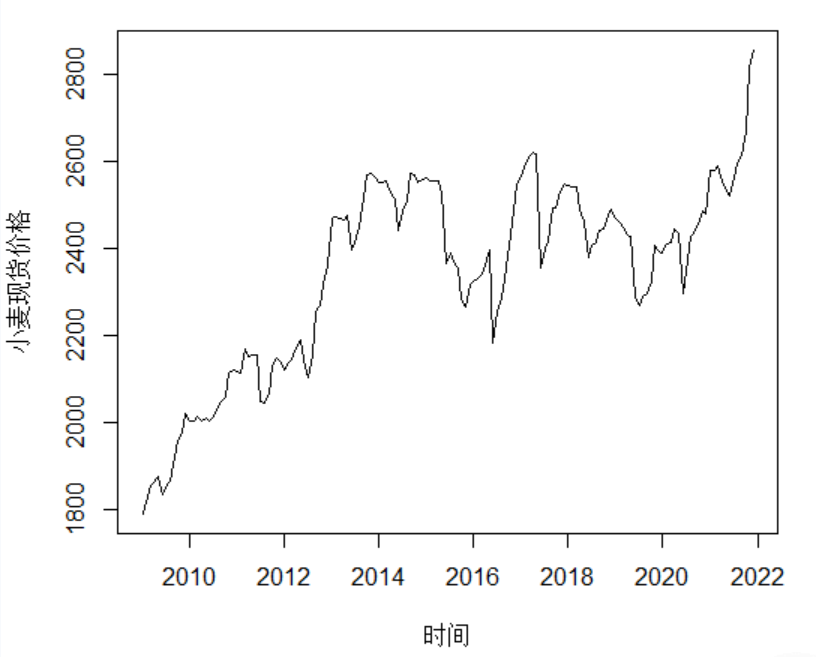


图 5 小麦现货价格原始数据时序图

从图5可以看出小麦现货价格呈现逐年递增的趋势，且具有季节性趋势，季节效应表现为12个月为一个周期，可以判断这个时间序列是非平稳的。由于季节效应的存在，选择模型时应考虑其效应的影响，首先排除非季节性的ARMA(p, q)模型。进一步运用单位根检验法对小麦现货价格时序的平稳性进行验证。由表4可以看出，在显著性水平为5%下，ADF检验的卡方统计量为-2.3799，且P-value值为0.4182（大于0.05），因此判定原序列存在单位根，为非平稳序列，需要先进行差分处理。

表 4 小麦现货价格时间序列的单位根检验

|  |  |  |
| --- | --- | --- |
|  | Dickey-Fuller | P-value |
| ADF test statistic | -2.3799 | 0.4182 |

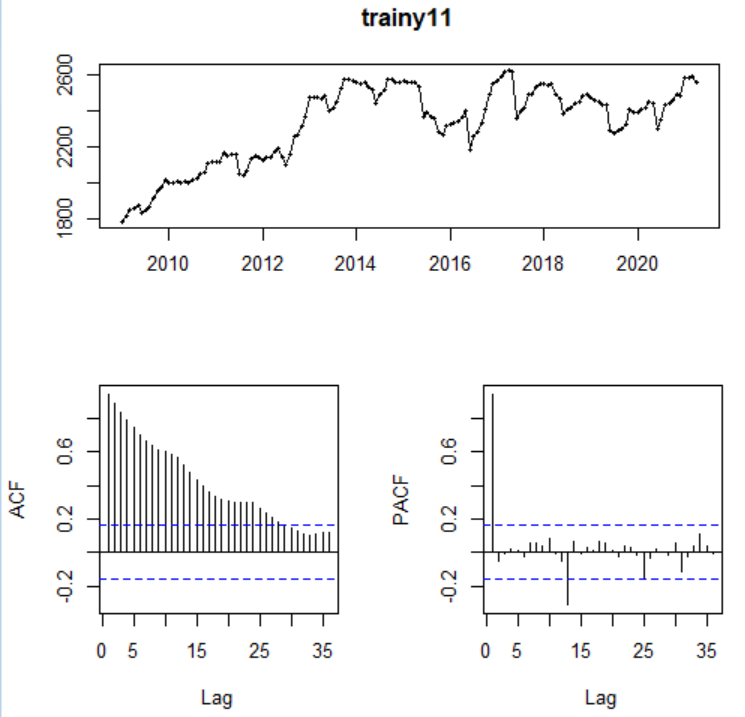


图 6 小麦现货价格原始数据自相关图与偏自相关图

表 5 ARMA模型定阶原则

|  |  |  |  |
| --- | --- | --- | --- |
| 模型定阶 | AR(p) | MR(q) | ARMA(p,q) |
| 自相关系数(ρk) | 拖尾 | P阶截尾 | 拖尾 |
| 偏自相关系数(φkk) | Q阶截尾 | 拖尾 | 拖尾 |

由小麦现货价格原始数据自相关图与偏自相关图可看出，ACF图中显示是指数衰减趋势，即拖尾现象，并且衰减速度缓慢，一直没有衰减到接近0的数值，仍有大量值大于0.5，说明序列yt与之前的时刻有强烈的相关性，而且相关项的范围是比较大的，可进一步可判断为非平稳时间序列；而在PACF图呈现一阶截断，显示为截尾，因此运用ARIMA(p, d, q)模型建模。

### 时间序列的平稳化操作

由于原时序数据为非平稳序列，为了消除线性趋势效应，我们取每一项与前一项的差：首先需要先对数据进行差分处理，将连续观察值之间做差，再对一阶差分后的数据做出时序图、ACF与PACF图检验。

从上图可见一阶差分后的时序图在0附近波动，但仍然存在较为显著的波动起伏大的情况，因此仍然不够平稳；且ACF图的滞后12期、滞后24期、滞后36期处显示序列存在强自相关，因此可以推测存在一个12阶的周期性影响因素。为了消除趋势效应，我们首先通过再次进行差分做出时序图、ACF与PACF图检验与单位根检验，判断差分后的序列是否平稳。

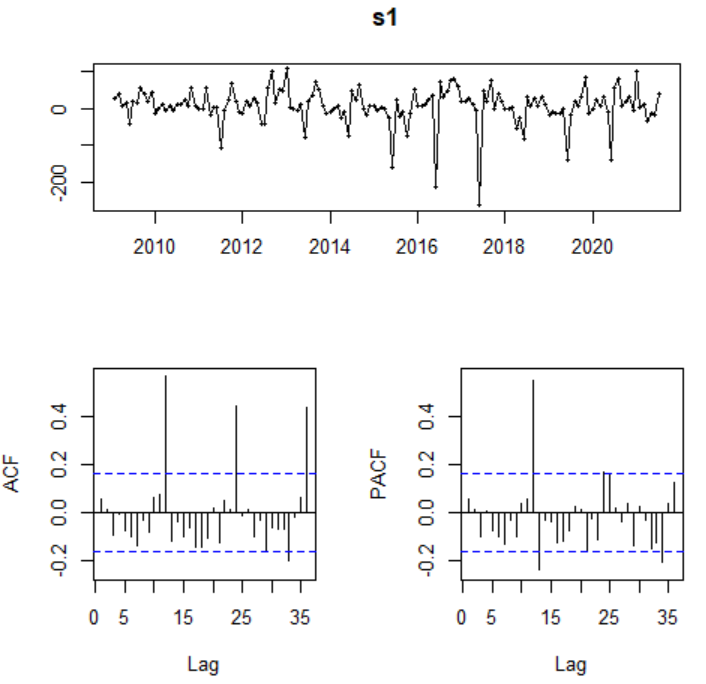


图 7 一阶差分和季节差分后的小麦现货价格时序图、ACF图和PACF图

表 6 季节差分后的小麦现货价格单位根检验

|  |  |  |
| --- | --- | --- |
|  | Dickey-Fuller | P-value |
| ADF test statistic | -5.6747 | 0.01 |

从上图可见一阶差分后的时序图在0附近波动，2015年后的差分序列存在较为显著的波动起伏大的情况，由ADF检验可以看出，在5%的水平下，卡方统计量值为-5.6747，且P值大于0.05，因此可以拒绝原假设，认为经过一阶差分的序列是平稳序列，符合ARMA时间序列模型的建立。由ACF图可以看出，在已经消除了单位根的情况下，滞后12期、滞后24期、滞后36期处仍然存在序列强自相关，因此考虑建立ARIMA(12,1,0)。

## 模型的优化选择与参数估计

下列为假设的ARIMA模型，根据最小信息量准则可以得到最优的模型。

表 7 ARIMA模型检验

|  |  |  |
| --- | --- | --- |
|  | 似然比统计量 | AIC |
| ARIMA((1,2,12),1,0) | -754.35 | 1516.7 |
| ARIMA((1,2,3,12),1,0) | -754.18 | 1518.35 |
| ARIMA((1,2,6,12),1,0) | -754.28 | 1518.55 |
| ARIMA((1,2,3,12),1,0) | -750.95 | 1511.91 |

根据简练原则，选取最小AIC与似然比统计量对应的模型，可得ARIMA((1,2,12),1,0)是最优的模型，以下是该模型的参数估计,可见系数在1%的显著性水平下显著，因此模型参数是有意义的。

表 8 ARIMA((1,2,12),1,0)模型系数表

|  |  |  |  |
| --- | --- | --- | --- |
|  | ar1 | ar2 | ar12 |
| Coeff | 0.0224 | -0.0111 | 0.5916 |
| 标准差 | 0.0650 | 0.0647 | 0.0685 |

由系数表可得最后估计出来的模型为：

式（1）

## ARIMA模型的显著性（白噪声）检验

为检验模型的有效性，采取观察残差标准差图、Ljung-Box法来检验模型的显著性。若上述各种检验方法不能拒绝原假设，则可以认为ARIMA((1,2,12),1,0)模型的残差序列是白噪声序列，该模型有意义。

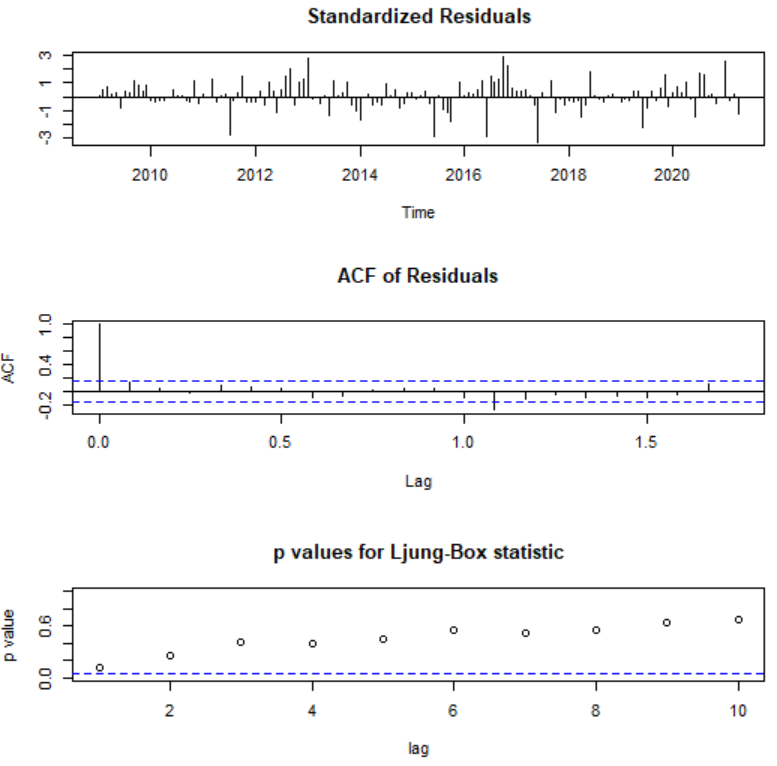


图 8 残差标准差、ACF图、Ljung-Box检验图

表 9 Ljung-Box检验

|  |  |  |
| --- | --- | --- |
|  | X-squared | p-value |
| model$residuals | 8.9772 | 0.7049 |

图8为标准化后所得到的残差图，可见残余误差都在零均值附近范围内波动，具有均匀的方差，没有明显的异常值。残差标准差的自回归函数都为近似0（两虚线内），Ljung-Box检验图中随着滞后阶数的增大，p值增加并总体在0.5上方，因此可以判断残差序列大致是白噪声序列。LB检验的p值为0.7049（大于0.05），不能拒绝原假设H0，可以认为该序列为白噪声序列，拟合模型ARIMA((1,2,12),1,0)为显著有效的模型。

## ARIMA模型的预测效果

利用训练好的模型对未来我国小麦的月度价格进行预测, 得到2021年5月到2021年9月的月度价格预测值。

表 10 2021年5月到2021年9月价格预测值

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 月份 | 真实值 | 预测值 | 误差 | 相对误差 | 平均相对误差 |
| 2021-05 | 2540.00 | 2549.449 | 9.45 | 0.37% | 1.89% |
| 2021-06 | 2522.22 | 2466.864 | -55.36 | -2.19% |
| 2021-07 | 2561.67 | 2496.633 | -65.04 | -2.54% |
| 2021-08 | 2595.56 | 2545.214 | -50.35 | -1.94% |
| 2021-09 | 2612.78 | 2549.592 | -63.19 | -2.42% |

运用均方根误差（RMSE）与平均绝对百分比误差（MAPE）两个模型预测的评价指标对小麦现货价格的预测结果进行评估。该AIRIMA模型的RMSE较小，平均绝对百分比误差MAPE仅为1.0347%，结果显示该模型拟合效果较好。

表 11 ARIMA模型检验

|  |  |
| --- | --- |
| 评价标准 | 结果 |
| MAE | 28.3136 |
| RMSE | 40.0983 |
| MAPE | 1.2230 |

根据以上ARIMA价格预测图以及预测结果可以看出，ARIMA模型能较准确地预测小麦现货价格的时间序列整体趋势。

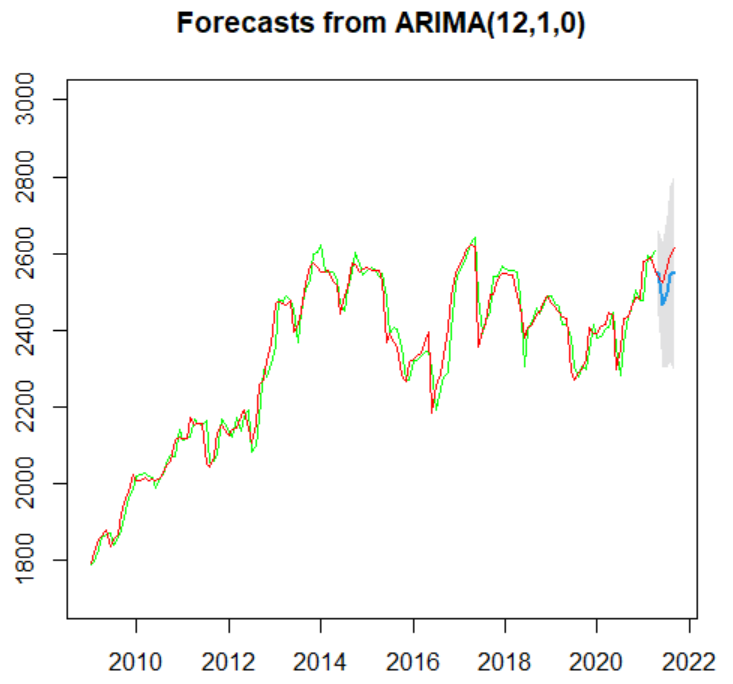


图 9 ARIMA模型预测图

# BP神经网络模型小麦价格预测

根据小麦价格序列不平稳的问题，以下考虑使用神经网络模型。神经网络算法众多，其中 BP 神经网络因其网络结构简单直观的特点而被广泛应用。

## BP神经网络理论概述

BP神经网络的概念由Rumelhart和McCelland在20世纪末提出，是利用误差逆传播方法在人工网络进行训练，被广泛应用于多因素、非线性和不确定的预测和评价问题。通过网络的变结构调节过程来学习训练样本，从而在不确定的环境中对复杂问题进行判断。

从结果上划分，BP神经网络包括输入层，隐含层和输出层，隐含层的数量为1个或者多个，输入层和输出层数量皆为1。层级之间的神经元相互连接，构成了神经网络基本结构。正向传播数据传递方向为输入层至隐含层至输出层。将样本数据代入网络进行训练和学习，使输出值不断接近期望值，不断调整权值和阈值，探索变量之间的关系，当误差达到可接受范围，学习过程就会结束；如果输出值误差超出预期范围，则进入误差反向传播，神经网络能够视误差的大小，由后至前逐层进行权重修正，从而降低误差，并重复上述过程，直至误差降低至预期范围。

### BP神经网络模型的构建

#### （1）网络层次确定

具有隐藏层的BP神经网络，所有闭合区间的连续函数去可以通过学习来进行模拟，寻找到数据之间的函数关系，从而提高映射的准确度，然而，这也会使得学习的周期加长，并且可能出现过拟合的现象。基于此，本章选择建立2层的BP神经网络模型。

#### （2）网络各层级点数确定

输入层和输出层节点数通常根据具体问题来决定。输入层节点的个数往往与输入的维数相同；输出层的变量为因变量，本文旨在对小麦现货价格进行预测,因此本文构建的神经网络模型的输出节点数量为1。

BP神经网络的隐藏层节点数的选取是一个比较复杂的问题，目前尚且没有统一的方法，往往根据经验和多次的试验来确定,不存在一个理想的解析式表示，因此本文参考目前常用的三个公式选择隐藏层节点数。

 式（2）

其中，k为样本数，为隐藏层节点数，为输入层节点数。如果，

 式（3）

其中代表输出节点数,代表输入层节点数,为之间的常数。

 式（4）

其中为输入层节点数。

因此本文隐藏层节点数在[3,14]之间。

### 模型训练

本次研究的BP神经网络模型的实证部分基于R语言中neuralnet程序包完成。首先将数据进行归一化处理，由于本文变量均为极大化指标，归一化公式为：

式（5）

进行BP神经网络模型预测后，对得到的训练集预测值进行反归一化以得到真正的预测数据。

## 单变量BP神经网络模型建模预测

经试验将隐藏层数定为2，两个隐藏层的神经元个数分别为5和4，目标收敛误差设置为0.01，使用 MSE(均方误差)作为损失函数，tanh 作为激活函数，最大迭代次数为 10000次，对归一化后的训练集进行建立2层的前馈神经网络模型。

本次研究将数据分成2部分，2009年1月至2021年4月的数据作为训练数据来训练神经网络模型，2021年5月至2021年9月的数据作为测试集用来检验神经网络模型。参考相关文献及前文ACF检验中推测小麦价格很可能存在一个滞后12期的影响因素（即小麦价格受到一年前同样月份的价格影响），本次研究采取将小麦价格滞后1、2、3、12期作为自变量进行单变量BP神经网络模型的建模预测。运行代码，结果如下：

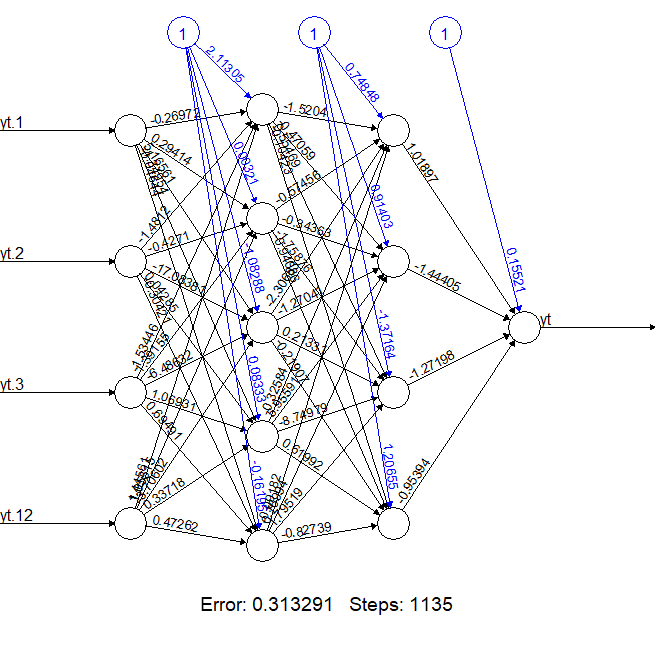


图 10 单变量BP神经网络模型

均方根误差可以衡量模型的预测性能。对于单变量BP神经网络模型，训练集的均方根误差为45.5，绘制训练集的拟合情况如下图所示：

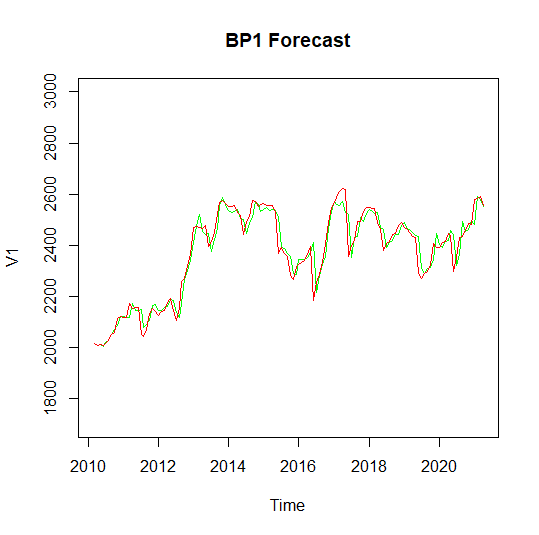


图 11 单变量BP神经网络模型训练集预测拟合情况

根据训练集拟合情况图，可以看出数据拟合程度较好。ARIMA模型的训练集均方根误差为28.36，与ARIMA模型相比，单变量BP模型对训练集的拟合效果不如ARIMA模型。

用测试集对训练集拟合的模型进行验证，得到测试集的均方根误差为40.9，比测试集略大。其详细预测结果见下表：

表 12 单变量BP神经网络模型测试集预测效果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 真实值 | 预测值 | 误差 | 相对误差 | 平均相对误差 |
| 2021-05 | 2540.00 | 2494 | -46.00 | -1.81% | 1.51% |
| 2021-06 | 2522.22 | 2502 | -20.22 | -0.80% |
| 2021-07 | 2561.67 | 2507 | -54.67 | -2.13% |
| 2021-08 | 2595.56 | 2551 | -44.56 | -1.72% |
| 2021-09 | 2612.78 | 2584 | -28.78 | -1.10% |

结果显示，单变量BP神经网络的预测值与实际值相比偏低，且误差逐渐变大。ARIMA模型的未来5期预测值平均相对误差为1.89%，从预测精度来看，单变量BP模型预测结果比ARIMA模型预测结果更好。

## 多变量BP神经网络模型建模预测

ARIMA模型在外推时不方便加入外生变量，即没有考虑结构变动带来的影响，如政策因素、新冠疫情等突发情况等，而我国小麦现货价格是政策、市场等多方面因素综合影响下的结果，因此用ARIMA模型可能导致未来的预测效果不好。因此此处考虑建立多变量的BP神经网络模型。

参考多变量时间序列的BP神经网络模型的建模方式，本节将第二章得到的6个影响因素引入，建立多变量的BP神经网络模型对小麦价格进行预测。经试验将隐藏层数定为2，两个隐藏层的神经元个数分别为5和4，目标收敛误差设置为0.01，使用 MSE(均方误差)作为损失函数，tanh 作为激活函数，最大迭代次数为 10000次，对归一化后的训练集进行建立2层的前馈神经网络模型。

本次研究将数据分成2部分，2009年1月至2021年4月的数据作为训练数据来训练神经网络模型，2021年5月至2021年9月的数据作为测试集用来检验神经网络模型。参考相关文献及前文ACF检验中推测小麦价格很可能存在一个滞后12期的影响因素（即小麦价格受到一年前同样月份的价格影响），本次研究采取尝试分别将小麦价格滞后1、2、3、12期作为自变量加入进行多变量BP神经网络模型的建模并取其中均方根误差最小的一个模型进行小麦价格的预测。

表 13 不同滞后期数的多变量BP模型预测效果对比

|  |  |  |
| --- | --- | --- |
|  | 均方根误差（训练集） | 均方根误差（测试集） |
| 滞后1、12期 | 37.8 | 34.2 |
| 滞后1、2、12期 | 34.2 | 16.7 |
| 滞后1、2、3、12期 | 39.4 | 98.8 |

根据上表可以看出，整体而言将小麦价格滞后1、2、12期加入自变量时，其训练集和测试集的模型均方根误差均为最小，预测效果最优。而将因变量滞后1、2、3、12期加入自变量时，其训练集均方根误差虽然也较小，但测试集预测效果呈现出不稳定的现象，向外预测的效果不佳。



图 12 多变量BP神经网络模型

绘制训练集的拟合情况如下图所示：

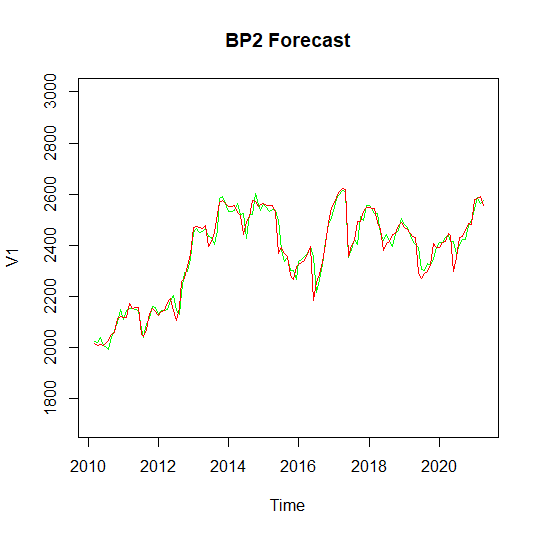


图 13 多变量BP神经网络模型训练集预测拟合情况

用测试集对训练集拟合的模型进行验证，预测结果见下表：

表 14 多变量BP神经网络模型测试集预测效果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 真实值 | 预测值 | 误差 | 相对误差 | 平均相对误差 |
| 2021-05 | 2540 | 2531 | -9 | -0.35% | 0.56% |
| 2021-06 | 2522 | 2498 | -24 | -0.95% |
| 2021-07 | 2562 | 2565 | 3 | 0.12% |
| 2021-08 | 2596 | 2621 | 25 | 0.96% |
| 2021-09 | 2613 | 2602 | -11 | -0.42% |

相比单变量BP神经网络模型，引入了影响因素后，多变量BP神经网络模型的预测效果明显提升，平均相对误差从1.51%降为 0.56%，且修正了传统ARIMA模型和单变量 BP模型预测值较真实值偏低的特点。从结果来看，BP神经网络在处理非线性序列上有着独特的优势。整体而言，多变量BP模型比ARIMA模型和单变量BP模型的预测效果更好，这启示我们在考虑问题时需要从整体出发，对影响问题的因素进行整体的探究，更有利于我们对问题进行全面的分析和推断。

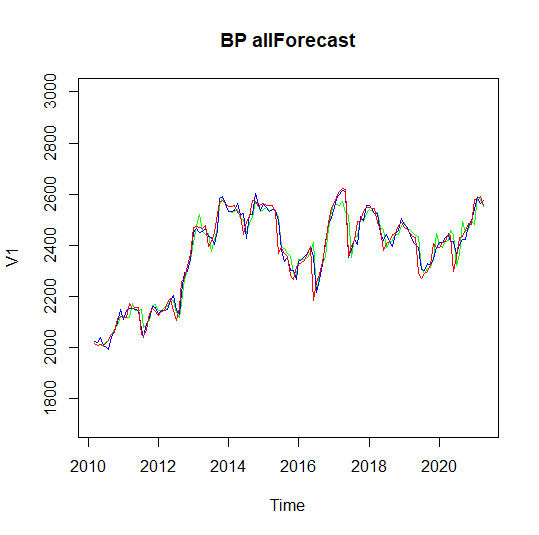


图 14 BP神经网络拟合结果

（红线：真实值 蓝线：多变量BP拟合值 绿线：单变量BP拟合值）

# ARIMA-BP组合预测模型

残差优化是一种“误差补偿”的思想，第一种模型得到的预测值与真实值的误差输入到第二种模型进行残差优化，使两种单一模型充分发挥各自优势，实现优势互补[27][28]。通过分析可知，小麦价格包含线性和非线性部分，本文选取对线性部分具有较高拟合度的 ARIMA 模型和对非线性部分解决效果较好的 BP 模型，并将其组合进行ARIMA -BP 残差优化组合模型预测分析，将小麦价格中的线性与非线性部分进行细化，充分提取数据信息，并应用于小麦价格预测研究，进行模型对比分析以探究方法的可行性。

ARIMA-BP 组合模型的构建思路如下[29]：将小麦价格序列分解为线性趋势和非线性趋势后，首先运用适用于预测线性趋势的 ARIMA 模型预测小麦价格线性部分，得到预测值 1；将真实值与预测值 1 作差得到残差值，将前 3 期的残差值输入到 BP 神经网络模型中对当期残差值进行预测，得到预测值 2，即为非线性趋势部分；将预测值 1 与预测值 2 加和，得到最终 ARIMA-BP 模型的预测值。

**(1) 线性趋势预测。**由第三章构建的 ARIMA 模型，得到2009年1月至2021年4月的原始小麦价格序列与 ARIMA 预测值的残差值如下图：

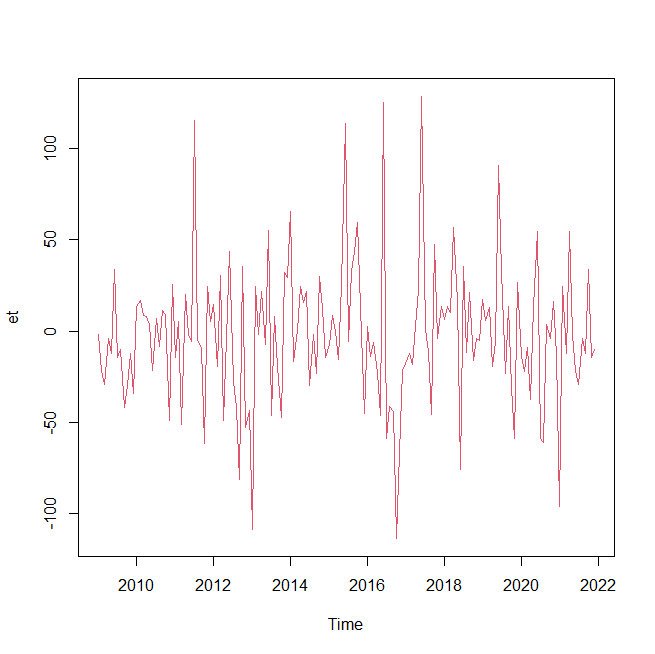


图 15 残差序列图

由残差序列图可以看出，残差值大的波动往往跟随着大的波动，小的波动往往跟随着小的波动，因此构造残差序列的自回归方程有利于帮助我们修正原ARIMA模型与真实值之间的差距。

**(2) 非线性趋势预测。**非线性趋势预测通过构建 BP 神经网络对(1)中的残差值预测实现。首先对残差部分作归一化处理，然后以滞后 1、2、3 期的残差值作为输入预测当期残差，得到2009年1月至2021年4月的残差数据的训练样本。

经试验将隐藏层数定为3，每个隐藏层的神经元个数为5，目标收敛误差设置为0.01，使用 MSE(均方误差)作为损失函数，tanh 作为激活函数，最大迭代次数为 10000次，对归一化后的训练集进行建立3层的前馈神经网络模型。



图 16 残差序列BP神经网络拟合模型

对得到的预测值进行反归一化处理，得到残差序列的真正预测值。将(1)和(2)中得到的预测值加和，即为 2009年1月至2021年4月训练集最终的预测值，使用2021年5月至2021年9月作为测试集数据进行验证，结果见下表：

表 15 ARIMA-BP组合模型测试集预测效果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 真实值 | ARIMA | 预测残差 | ARIMA-BP | 误差 | 相对误差 | 平均相对误差 |
| 2021-05 | 2540.00 | 2538.918 | -1.18 | 2540.10 | 0.10 | 0.00% | 1.25% |
| 2021-06 | 2522.22 | 2457.344 | -31.90 | 2489.24 | -32.98 | -1.31% |
| 2021-07 | 2561.67 | 2503.054 | -84.00 | 2587.05 | 25.38 | 0.99% |
| 2021-08 | 2595.56 | 2541.991 | -11.70 | 2553.69 | -41.87 | -1.61% |
| 2021-09 | 2612.78 | 2539.56 | -11.70 | 2551.26 | -61.52 | -2.35% |

相较于使用单一模型 ARIMA 模型进行预测，各年份误差、相对误差均有所下降，整体的预测效果更优，这主要应该是由于ARIMA模型预测结果往往较真实值偏小，而ARIMA-BP模型对此进行了一定程度的修正。

## 4.6模型比较

将本次研究建立的模型进行比较，结果如下表所示：

表 16 预测效果的模型结果比较

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 真实值 | ARIMA | 单变量BP | 多变量BP | ARIMA-BP |
| 2021-05 | 2540 | 2549 | 2494 | 2531 | 2540 |
| 2021-06 | 2522 | 2467 | 2502 | 2498 | 2489 |
| 2021-07 | 2562 | 2497 | 2507 | 2565 | 2587 |
| 2021-08 | 2596 | 2545 | 2551 | 2621 | 2554 |
| 2021-09 | 2613 | 2549 | 2584 | 2602 | 2551 |
| 平均相对误差 |  | 1.89% | 1.51% | 0.56% | 1.25% |

对比4个模型结果，发现：

（1）对于训练集样本，单变量BP模型不如ARIMA模型拟合效果好；对于测试集，单变量BP模型的外推预测效果比ARIMA精度稍高，这体现了BP神经网络在处理非线性序列的外推预测上有着独特的优势。

（2）考虑到加入影响因素能够有效提升模型的预测精度，根据第二章构建的小麦价格指标体系，构建多变量BP神经网络模型，经过验证，加入影响因素后小麦价格测试集的平均相对误差均降低，模型能够最为有效的预测小麦价格。

（3）本次研究尝试将文献提出的ARIMA-BP组合预测模型运用于小麦价格序列的建模预测。发现ARIMA-BP组合预测模型对于同样只使用了一个时间序列（即小麦现货价格时间序列）进行预测建模的的单变量序列的ARIMA 模型和单变量BP神经网络模型，其模型预测精度较高，稳定性更好。但ARIMA-BP组合预测模型的预测效果不及多变量BP神经网络模型。

# 5 项目特色与创新点

## 5.1选题具有现实意义

当前全球各地疫情连续不断，阻碍各国进行有规律的农业生产，而近日俄乌局势动荡，乌克兰作为欧洲粮仓的地位被打破，导致以小麦为代表的国际与国内粮食价格被不断推高，直接增加了老百姓的生活成本与市场风险，也使国家难以对小麦价格进行调控。而对小麦价格的准确预测有利于国家积极应对市场波动并采取有效的政策策略，从而降低市场风险。

## 5.2 研究方法具有创新性

本课题基于stacking集成学习算法中分类器的特点，创新性地将传统的arima模型以及LSTM、XGBoost模型组合成多变量的预测模型。目前XGBoost算法和LSTM算法在多个领域的预测问题上取得了较好的成果，但二者在经济预测方面的研究较少，结合机器学习方法在经济序列预测的较好成绩，考虑引用影响因素、使用组合模型进行预测能有效地提高模型预测效果。

（1）本文建立了有关小麦价格预测的指标体系，并利用相关系数与随机森林进行变量筛选，综合考虑了变量间的线性和非线性相关程度，筛选出对小麦现货价格变化贡献较大的特征，有利于提高模型的预测能力与拟合速度。

（2）本项目考虑到小麦现货价格同时具有线性与非线性特征，选择了适用于线性预测、短期预测精度较高的ARIMA模型拟合趋势，以及选择在处理非线性时间序列数据方面性能较好、泛化能力较强的BP神经网络模型对残差建模，建立的ARIMA-BP组合预测模型相比于单一预测模型具有提高精度的优势。

（3）在建立单变量和多变量BP神经网络模型时，综合考虑了小麦价格时间序列本身的自相关图确定滞后阶数，发现小麦价格很可能存在一个滞后12期的影响因素（即小麦价格受到一年前同样月份的价格影响），从而帮助我们确定BP神经网络模型的输入层节点数，对于构建BP神经网络有一定的指导意义。

（4）本文根据加法结构，分别建立ARIMA模型、单变量BP神经网络模型、多变量BP神经网络模型以及ARIMA-BP神经网络组合模型进行预测精度的比较，并得出了最优模型，丰富了有关市场价格预测的理论研究。

# 6 参考文献

[1]朱晶,张庆萍.中国利用俄罗斯、乌克兰和哈萨克斯坦小麦市场分析[J].农业经济问题,2014,35(04):42-50+111.DOI:10.13246/j.cnki.iae.2014.04.007.

[2]杨惠珍,韦敬楠,张立中.我国粮食期货市场价格发现功能的实证分析——以玉米和小麦市场为例[J].价格月刊,2017(05):19-23.DOI:10.14076/j.issn.1006-2025.2017.05.04.

[3]刘平,方旖旎.中国小麦进口结构及竞争力比较研究[J].市场周刊,2021,34(08):144-146.

[4]高云,孙一铮,郭新宇,矫健.国内外小麦价格相关性及影响因素研究[J].价格理论与实践,2018(05):71-74.DOI:10.19851/j.cnki.cn11-1010/f.2018.05.018.

[5]林春霞. 国际粮价将呈高位震荡趋势[N]. 中国经济时报,2021-09-29(001).DOI:10.28427/n.cnki.njjsb.2021.002172.

[6]赵荣,曹洁,朱婷婷.浅议中国小麦的国际贸易及竞争力提升[J].滁州学院学报,2021,23(01):14-19+41.

[7]朱聪,曲春红,王永春,赵伟.新一轮国际粮食价格上涨：原因及对中国市场的影响[J/OL].中国农业资源与区划:1-13[2022-03-08].http://kns.cnki.net/kcms/detail/11.3513.S.20211130.1225.010.html.

[8]钟钰,陈希,牛坤玉.粮食出口限制政策的实施效果与我国应对——来自部分小麦出口国的证据[J].经济纵横,2021(08):29-39+137.DOI:10.16528/j.cnki.22-1054/f.202108029.

[9]潘寅茹. 乌克兰为何被称为“欧洲粮仓”？玉米小麦出口均占全球10%以上[N]. 第一财经日报,2022-03-04(010).

[10]曹慧.新冠肺炎疫情对国内外小麦产业的冲击:趋势、问题与建议[J].世界农业,2021(01):4-10+27.DOI:10.13856/j.cn11-1097/s.2021.01.001.

[11]申洪源. 2021年中国小麦市场分析[N]. 粮油市场报,2022-02-15(B02).DOI:10.28553/n.cnki.nlysc.2022.000245.

[12]邵海鹏. 小麦也“疯狂”：进口首次突破配额，价格站上历史高位[N]. 第一财经日报,2022-03-02(A06).

[13]邵海鹏. CBOT小麦期货连续涨停，国内外小麦价格倒挂影响几何[N]. 第一财经日报,2022-03-04(A10).

[14]孙致陆.贸易开放背景下国际小麦贸易市场势力实证分析[J].华中农业大学学报(社会科学版),2019(04):1-14+169.DOI:10.13300/j.cnki.hnwkxb.2019.04.001.

[15]陈植. 芝商所小麦玉米期货迭创新高 对冲基金押宝“新避险投资”[N]. 21世纪经济报道,2022-03-04(005).

[16]Haile M G, Kalkuhl M. Braun J V Worldwide Acreage and Yield Response to International Price Change and Volatility: A D ynamic Panel Data Analysis for Wheat, Rice, Corn, and Soybeans M//Food Price Volatility and Its Implications for Food Secu rity and Policy. Springer International Publishing, 2016

[17]Pal D, Mitra S K Interdependence between crude oil and world food prices: A detrended cross correlation analysis. Phys ica A Statistical Mechanics&Its Applications, 2018

[18]Ahumada H, Cornejo M Forecasting food prices: The case of corn, soybeans and wheatly] International Journal of Foreca sting. 2016.

[19]柳燕子.基于多模型分析的小麦价格预测研究[J].山西农经,2018(04):148.DOI:10.16675/j.cnki.cn14-1065/f.2018.04.103.

[20]李雪,韩一军,付文阁.最低收购价政策对小麦市场价格波动影响的实证分析[J].华中农业大学学报(社会科学版),2018(02):1-7+154.DOI:10.13300/j.cnki.hnwkxb.2018.02.001.

[21]张瑞娟,高芸.国内外小麦价格联动关系研究[J].价格理论与实践 ,2016(07):112-115.DOI:10.19851/j.cnki.cn11-1010/f.2016.07.029.

[22]新冠肺炎疫情对中国粮食价格的影响效应及作用机制.崔海莹

[23] 战玉锋,刘放. 我国小麦期货价格影响因素实证分析[J]. 沈阳工业大学学报(社会科学版), 2012, 5(3): 223-231.

[24] 张恒,赵宇洋,安起光. 基于BEKK-GARCH模型的小麦期现货市场价格波动溢出性分析[J]. 粮食科技与经济, 2022, 47(1): 40-44.

[25]张鑫, 吴海涛, 曹雪虹. Hadoop 环境下基于随机森林的特征选择算法[J]. 计算机技术与发展, 2018, 28(07):88-92.

[26] 顾鑫. 基于ARIMA-LSTM-XGBoost组合模型的中国经济发展预测[D]. 青岛大学, 2021.

[27] GAO W J，LI Z，MIN X，et al. Application system response time pre⁃diction based on ARIMA-LSTM composite mode［J］. Computer &Digital Engineering，2021，49（5）：880-885.高文俊，李志，闵星，等 . 基于 ARIMA-LSTM 组合模型的应用系统响应时间预测［J］. 计算机与数字工程，2021，49（5）：880-885.

[28] XU D D，JIANG Z X. End-to-end speech recognition based on deep⁃ly optimized residuals convolutional neural network［J］. ApplicationResearch of Computers，2020，37（S2）：139-141.徐冬冬，蒋志翔 . 基于深度优化残差卷积神经网络的端到端语音识别［J］. 计算机应用研究，2020，37（S2）：139-141.

[29]Du Y. Predicting China GDP Based on ARIMA-BP Neural Network Model[A]. InformationEngineering Research Institute(USA) 、 Asia Pacific Human-Computer Interaction ResearchCenter(Hong Kong).Proceedings of 2018 IEA 3rd International Conference on Frontier ofComputer Science and Information Engineering(ICFCI 2018)[C].Information EngineeringResearch Institute(USA) 、 Asia Pacific Human-Computer Interaction Research Center(HongKong):智能信息技术应用学会, 2018:6.