基于 ARIMA-RF 组合模型的国内动力煤价格预测

扫码阅读

季凌雲，陆 伟，卓 辉，李佐健
（安徽理工大学 安全科学与工程学院，安徽 淮南 232001）

摘 要：选取秦皇岛港口动力煤价格作为研究对象，搜集 10 年间煤价数据并分析其影响因素，确定煤炭产量、港口库存、运输成本、火力发电量及社会用电量为主要影响因素；分别建立 ARIMA（2,1,2）模型和 RF（随机森林）模型并优化，通过加权平均法得到 ARIMA 和 RF 模型权重，建立 ARIMA-RF 组合模型。该模型较深度神经网络模型（DNN）、支持向量回归模型（SVR）、ARIMA 模型、RF 模型预测的煤价准确度更高，可准确预测动力煤价格走势，为调控能源消费强度、深化能源体制机制改革政策制定提供参考。

关键词：煤价预测；ARIMA 模型；随机森林模型；组合模型；精度优化
中图分类号：F416.21 **文献标志码**：A **文章编号**：1002-9605（2023）04-0028-10

Prediction of domestic thermal coal price based on ARIMA-RF combined model

Ji Lingyun, Lu Wei, Zhuo Hui, Li Zuojian
(*College of Safety Science and Engineering, Anhui University of Science and Technology, Huainan* 232001, *China*)

Abstract: This paper selects the thermal coal price of Qinhuangdao port as the research object, collects coal price data over the past decade and analyzes its influencing factors, the main influencing factors are determined to be coal production, port inventory, transportation costs, thermal power generation, and social electricity consumption. The ARIMA (2,1,2) model and RF model are established and optimized respectively. The weights of ARIMA and RF models are obtained through the weighted average method and the combined model of ARIMA and RF (Random Forest) is established. Compared to deep neural network models (DNN), support vector regression models (SVR), ARIMA models, and RF models, this model has higher accuracy in predicting coal prices. This model can accurately predict the thermal coal price, proving reference for regulating energy consumption intensity, deepening energy system and mechanism reform, and policy-making.

Key words: coal price prediction; ARIMA model; Random Forest model; combined model; accuracy optimization

0 引 言

实现碳达峰、碳中和，是以习近平同志为核心的党中央统筹国内国际两个大局作出的重大战略决策；是着力解决资源环境约束突出问题，实现中华民族永续发展的必然选择[1]。而“推动经济社会发展全面绿色转型，加快构建清洁低碳安全高效能源体系”是“双碳”工作重点任务。在此过程中，需要“强化能源消费强度和总量双控，深化能源体制机制改革”[2]。动力煤价格作为能源领域的重要指标，既切实反映了我国能源绿色低碳转型、产业结构深度调整情况，又和全球煤炭市场供需状态及国内经济社会稳定发展息息相关[3]。因此，动力煤价格的准确预测对确保能源安全稳定供应和平稳过渡，保障煤炭生产、流通、消费等环节平稳运行具有重要意义。

国内外学者从不同角度，选用不同的研究方法对煤价进行预测，目前常用的方法有灰色模型（GM）、神经网络模型、经验模态分解（EMD）、时间序列模型和组合模型[4-8]。王帮俊等[9]选用相空间重构技术，对煤炭价格的时间序列进行分析，发现国内外煤炭价格的有效预测周期分别为 12 个月和 5 个月。向超[10]采用 SVR 和 ARIMA 相组合的模型，对煤价进行了预测，对比各个组合模型的预测精度，发现变权组合模型的预测结果优于平权组合模型和单一预测模型。Pan Sitong[11] 以日、周、

基金项目：安徽省高校优秀科研创新团队项目（2022AH010051）；国家自然科学基金（51974178）；安徽省重点研究与开发计划项目（2022m07020006）

通讯作者：卓 辉（1992—），男，安徽宿州人，讲师，研究方向为矿井火灾防治、煤自燃、数值模拟、热动力灾害监测预警等。E-mail：zhuohui1130@126.com

月为时间单位，建立长短期记忆递归神经网络模型（LSTM），预测煤炭价格变化趋势，经过精度和灵敏度检验，表明该模型稳定。Xu Xiaojie 等[12]建立非线性自回归神经网络来研究焦煤期货的日收盘价格，验证了机器学习技术对焦煤价格预测的有效性。Matyjaszek 等[13]比较了传统时间序列模型、ROBUST 模型、ARIMA 模型、广义回归神经网络（GANNs）和多层前馈网络（MLFNs）对焦煤价格的预测性能，发现 ARIMA 模型在转基因时间序列预测时性能更高。Alameer 等[14]结合长短期记忆（LSTM）和深度神经网络（DNN），准确预测了不同层位的月度煤价波动。Yang Shaomei 等[15]提出基于集成经验模态分解（MEEMD）和改进的鲸鱼优化算法（IWOA）优化的 LSTM 混合模型，用于预测煤炭期货价格比其他 11 个模型性能更好。Fu Xiangwan 等[16]使用 EMD 和 ARIMA 组合模型预测中国动力煤价格，对比发现其评价指标数值更优。上述模型的主要目的是采用预测结果更为精确的模型预测煤价，但未对煤价波动的影响因素进行深入分析[17]。

Liu Xinrong[18]选取影响煤炭价格的 6 个因素，提出一种改进的 Adam 优化器 LSTM 神经网络模型预测煤价，表明其预测性能优于传统模型。Ding Lili 等[19]基于可再生能源、大庆油田、日本天然气、澳大利亚蒸汽煤价格、煤矿行业指数、A 股电力部门指数、A 股指数、煤炭行业指数等日因素，对我国每周汽煤价格进行概率密度预测，发现澳大利亚汽煤价格、可再生能源和 A 股指数是汽煤价格的 3 个最佳预测因子。Wang Xiaofei 等[20]采用双重差分模型定量评估煤炭去产能政策对煤炭价格的影响效果，得出两项去产能政策的及时性存在显著差异。Zhu Shiqiu 等[21]采用协整检验和预测误差方差分解（FEVD）分析了影响煤价的因素，发现政策不确定性对煤价影响为负且弱，而能源价格对煤价的正向影响更大。Li Yanbin 等[22]采用灰色关联分析法构建煤炭需求量影响因素体系，引入粒子群算法（PSO），建立了改进的 GSA（IGSA）-SVM 预测模型。Zhang Lihui 等[23]基于鲁棒主成分分析（RPCA），得到 5 个影响能源消费的主要因素，比较并证明了所提出的 TS-PSO-LSSVM 模型预测精度更高、训练速度更快。廖志伟等[24]基于卡方分析和相关系数筛选中短期煤价的主要影响因素，采用 LSTM 神经网络，验证了此深度学习模型的有效性。Sohrabi 等[25]考虑经济和政治条件，将平均收益布朗运动（BMMR）和 RBF 神经网络模型（CBRN）结合，得到误差较低的煤价预测模型。Feng Xiwen 等[26]利用协整技术，从能源和经济结构角度研究中国煤炭的长期关系，发现经济结构调整抑制煤炭需求，生产者价格的变动对煤炭需求的变动影响不大，天然气消费对煤炭需求的影响越来越大。Gao Luhui 等[27]采用灰色关联分析法描述主要因素与煤价的关联度，根据相关系数对主要影响因素排序，借助协整理论改进传统神经网络，得到有效的煤价预测模型。上述模型的预测精度进一步优化，但没有充分考虑煤价的非线性波动特征和滞后性对煤价预测的影响，需要继续对煤价影响因素分析降维，构建更具鲁棒性、泛化性和高精度的预测模型。

本文基于煤价的供需因素，依据不同时间频度的数据，提出一种基于 ARIMA 和 RF 组合的国内动力煤价格预测模型，力求实现我国动力煤价格的准确预测。首先，分析煤价的供需关系，依据相关性分析验证煤价影响因素的可靠性，采用均值插补法对数据补缺，将不同粒度的数据通过 EViews 转换成周度数据，通过因子分析降低维度。其次，基于贝叶斯最优准则（BIC）及 Ljung-Box（LB）检验，建立预测模型 ARIMA（2,1,2）。然后采用网格搜索法优化基于决策树的 RF 模型的参数，通过加权平均法并联 ARIMA 和 RF 模型，组成优化后的 ARIMA-RF 组合模型。最后根据多种模型的对比分析，验证 ARIMA-RF 组合模型的准确性与精确性。该模型可为政府制定相关政策引导煤炭市场健康发展提供一定参考，有助于能源安全稳定供应和平稳过渡及“双碳”目标的早日实现。

1 数据预处理

1.1 数据选取

动力煤在不同的港口往往价格差距甚大，为了能公平客观地体现市场面上的动力煤价，各个相关机构一般使用港口平仓价格作为动力煤的价格指标。秦皇岛港处于环渤海经济圈中，是环渤海十大港口之一，也是国家唯一直接进行管理的港口；同时作为煤炭输出港口，每年的煤炭吞吐量超过 2 亿 t。本文选取秦皇岛港口动力煤价格作为研究对象，对其近 10 年动力煤价的形成进行解析。我国煤炭行业在运输方面基本形成“西煤东送，北煤南输”的总体格局。煤炭从西部产出后，通过铁路或公路运输集中到沿海港口，再通过装船从海上运送到长江三角

洲和珠江三角洲，之后通过陆路交通运输到各个火力发电厂。通过对其中的供需关系进行分析，初步提炼出影响煤价的基本指标，见表 1。

表 1 影响煤价的基本指标

基本指标	数据来源
煤炭产量	内蒙古原煤产量（月）
港口库存	秦皇岛港煤炭库存合计（日）
运输成本	CBCFI 中国沿海煤炭运价指数（周）
火力发电量	各省市火电机组当月发电量（全国）（月）
社会用电量	全社会用电量总计（月）
动力煤价格	CCTD 秦皇岛动力煤平仓价格（日）

1.2 数据清洗

数据补插方法有补插均值、中位数、众数，使用固定值，最近邻补插，回归方法和数学插值法。本文使用的数据完整性较好，且通过对比，采用均值插补法最佳。不同指标数据时间频度有较大差异，本文使用 EViews 软件进行数据转频。将选定的 2011—2022 年的煤炭产量（月）、火力发电量（月）和社会用电量（月）所有数据以 Sum 为基准的二次插值，转为 2011—2022 年的周度数据。将 2011—2022 年的港口库存（日）和动力煤价格（日）以 Average 为基准的二次插值，同样转成同频的周度数据。

1.3 相关性分析

首先对完成数据清洗后的数据进行相关性分析，初步了解数据间的线性关系，为后文的分析提供依据。本文使用皮尔森相关系数对数据进行分析，X、Y 为需要进行相关性分析的数据对象，x_i、y_i 为这 2 个对象的取值，见下式。

$$\rho_{xy}=\frac{cov(X,\ Y)}{\sqrt{Var(X)\ Var(Y)}}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\ (y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{1}$$

通过 Python 运算后得出的效果图如图 1 所示。

由图 1 初步分析可知，动力煤的价格与运输成本、火力发电量、社会用电量和煤炭产量呈现正相关，且与运输成本关联性最大，呈强相关态；动力煤的价格与港口库存呈现负相关。这与初步分析基本一致，验证表明数据选取相对正确，具有可靠性。

1.4 因子分析

因子分析是一种常用的统计分析方法，通常采

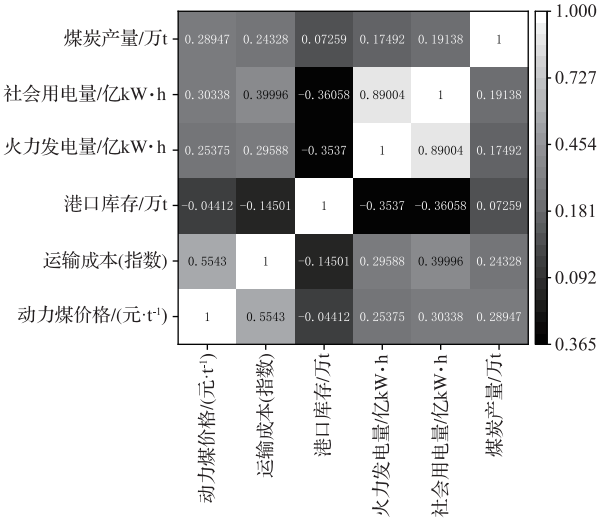

图 1 影响因素相关性分析

用 SPSS 软件进行统计分析。它通过对变量之间的相关系数矩阵进行计算，并依据数据之间的相关性对变量进行分组，使强相关的数据分为一组，弱相关或不相关的数据分到不同组；而后选举出代表每组中数据基本趋势的新变量作为公共因子，将多个指标聚合成两三个公共因子，从而达到降低变量维度的目的。

第 1 步是检验 KMO 和 Bartlett 数值，以此判断因子分析的适用性。KMO 值在 0.9 以上，说明因子分析的适用性极强；在 0.7~0.9 之间表示强；0.6~0.7 之间适中；0.5~0.6 之间表示弱；0.5 以下极弱。将动力煤价格及煤价影响因素数据导入 SPSS 软件中进行检验，结果见表 2。

表 2 KMO 检验和 Bartlett 检验

KMO 值	Bartlett 球度检验		
	近似卡方	df	p
0.605	1 067.171	10.000	0.000 0

可以看出，KMO 统计量为 0.605，在程度上，本数据进行后续分析具有合理性。Bartlett 球度检验的显著性为 1%，水平上呈现显著性，拒绝原假设，各变量间具有相关性，因子分析有效。

通过分析总方差解释表和旋转后因子载荷系数表，查看因子数和对数据解释效果。总方差解释表主要是看因子对于变量解释程度的贡献率，若贡献率太低，说明总体解释效果差，则需要调整因子数量；旋转后因子载荷系数表的作用是查看每个因子对所有输入数据的解释效果，数值越高表示本因子对当前变量解释效果越好；共同度是所有因子对当

前变量的综合解释度，数值与解释效果呈现正相关。利用 SPSS 软件分析得总方差解释表和旋转后因子载荷系数，见表 3 和表 4。

表 3 总方差解释

成分	特征根			旋转后方差解释率		
	特征根	方差百分比/%	累积/%	特征根	方差百分比/%	累积/%
1	2.375	47.503	47.503	2.062	41.247	41.247
2	1.124	22.483	69.986	1.105	22.109	63.357
3	0.734	14.676	84.662	1.065	21.305	84.662
4	0.664	13.282	97.944	—	—	—
5	0.103	2.056	100.0	—	—	—

表 4 旋转后因子载荷系数

项目	旋转后因子载荷系数			共同度（公因子方差）
	因子 1	因子 2	因子 3	
煤炭产量/万 t	0.225	0.267	0.805	0.770
运输成本/指数	0.170	0.952	0.081	0.942
港口库存/万 t	−0.444	−0.269	0.640	0.680
火力发电量/（亿 kW·h）	0.958	0.092	0.013	0.927
社会用电量/（亿 kW·h）	0.931	0.215	0.013	0.914

在方差解释表中，当主成分为 3 时，总方差解释的特征根低于 1.0，总方差解释率达到 84.662%，对原始数据有着较为不错的总体解释效果。旋转后因子载荷系数表中，3 个因子分别对 5 组原始数据的公因子方差最低为 0.680，处于 0.600 标准以上，对所有单个数据都有着较好的解释水平。验证表明公共因子的选择较为成功。

最后通过分析成分矩阵，利用提取到的 3 个公共因子对原始数据进行定义。成分矩阵见表 5。

表 5 成分矩阵

项目	成分		
	成分 1	成分 2	成分 3
煤炭产量/万 t	0.095	0.237	1.097
运输成本/指数	0.071	0.847	0.11
港口库存/万 t	−0.187	−0.239	0.873
火力发电量/（亿 kW·h）	0.403	0.082	0.018
社会用电量/（亿 kW·h）	0.392	0.192	0.018

利用上表计算出成分得分，得出主成分公式如下。

$$\begin{cases} F_1 = 0.095 \times 煤炭产量 + 0.071 \times 运输成本 - 0.187 \times 港口库存 + 0.403 \times 火力发电量 + 0.392 \times 社会用电量 \\ F_2 = 0.237 \times 煤炭产量 + 0.847 \times 运输成本 - 0.239 \times 港口库存 + 0.082 \times 火力发电量 + 0.192 \times 社会用电量 \\ F_3 = 1.097 \times 煤炭产量 + 0.11 \times 运输成本 + 0.873 \times 港口库存 + 0.018 \times 火力发电量 + 0.018 \times 社会用电量 \\ F = \dfrac{0.412}{0.847} \times F_1 + \dfrac{0.221}{0.847} \times F_2 + \dfrac{0.213}{0.847} \times F_3 \end{cases} \tag{2}$$

2 模型建立

2.1 ARIMA 模型

ARIMA 模型即自回归积分滑动平均模型，是一种通过历史时间序列数据预测未来一段时间目标数据的预测方法，由自回归滑动平均（ARMA）模型扩展而来。时间序列模型主要包括 AR（p）模型（自回归），p 为自回归移动阶数；MA（q）模型（滑动平均），q 为移动平均阶数；ARMA（p，q）模型（自回归滑动平均）；ARIMA（p,d,q）模型（自回归积分滑动平均），d 为差分阶数。

AR（p）模型，如果时间序列 $\{X_t\}$ 满足下式，则说明其是自回归移动阶数为 p 阶的自回归模型。

$$X_t = \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + \beta_p X_{t-p} + \varepsilon_t \tag{3}$$

其中，X_t 为待预测的期望值，X_{t-o} 为之前 o 期的值，对应的 β_o 为该数值的系数；ε_t 为 1 个独立同分布的随机变量序列，其均值为零，方差大于零，即纯随机序列，又称为白噪声序列。

MA（q）模型，如果序列 $\{X_t\}$ 满足下式，说明其是移动平均阶数为 q 阶的滑动平均模型。

$$X_t = \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \alpha_2 \varepsilon_{t-2} + \cdots + \alpha_q \varepsilon_{t-q} \tag{4}$$

其中，X_t 意义同上；ε_{t-o} 为过去 o 期对应的随机干扰值，对应 α_o 为此数据的系数。

ARMA（p，q）模型，此自回归滑动平均过程中的序列 $\{X_t\}$，满足下式。

$$X_t = \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + \beta_p X_{t-p} + \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \alpha_2 \varepsilon_{t-2} + \cdots + \alpha_q \varepsilon_{t-q} \tag{5}$$

其中，X_t、X_{t-i}、β_i、ε_{t-i}、α_i 和 AR（p）、MA（q）模型中的含义相同。

以上 3 个时间序列模型都是基于平稳的时间序列。当时间序列因为具有趋势性而不平稳后，需要对

其进行必要的 d 次差分，以消除趋势影响，使之平稳；然后可使用 ARMA(p,q)模型，进行分析预测。

ARIMA（p,d,q）模型，自回归积分滑动平均过程的非平稳序列 $\{Y_t\}$ 满足下式。

$$\begin{cases}\varphi(B)\ \nabla^d X_t = (1-\alpha_1 B-\alpha_2 B^2-\cdots-\alpha_q B^q)\ \varepsilon_t,\\ E(\varepsilon_t)=0,\ Var(\varepsilon_t)=\sigma_s^2,\ E(\varepsilon_s\varepsilon_t)=0,\ s\neq t,\\ E(X_s\varepsilon_t)=0,\ \forall s<t\end{cases} \tag{6}$$

其中，$\nabla^d=(1-B)^2$，B 是后移算子，∇ 为差分符号，σ_s 是纯随机序列 ε_t 的方差，其他符号含义与 AR（p）、MA（q）、ARMA（p，q）模型的含义相同。

ARIMA 模型的算法实现过程，如图 2 所示。

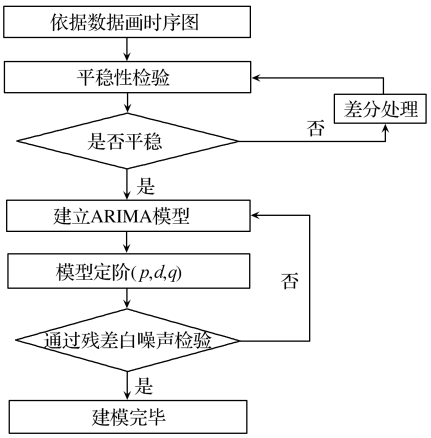

图 2 ARIMA 建模流程

2.2 RF 模型

随机森林（RF）算法是为了解决单一决策树模型存在的缺陷，将多个决策树模型进行叠加，形成“森林”，以此预测最终的结果。它的输出是若干决策树输出的集合。决策树算法是一种经典的机器学习算法，其本质是数据结构中的二叉树，开始只有根节点，通过不断划分生成新的节点，连接节点之间的线为有向边，叶节点是决策树中最顶端的节点，决策树唯一的输出是从根节点到某个叶节点的路径的值。决策树需要使用预处理过的数据作为训练样本进行构建。节点划分的依据是节点的“纯度”，即划分后的节点“纯度”要高于划分前，否则将不对此叶节点进行划分。一般采取“基尼值”（Gini）来表示节点样本数据间的纯度，计算过程如下。

1）样本 Gini 指标计算。

$$Gini(S)=1-\sum_{i=1}^{n}P_i^2 \tag{7}$$

其中，S 为数据集合，n 为样本类别数目，P_i 为 S 中第 i 类样本所占的比例。$Gini$（S）表示从数据集 S 中随机选择 2 个样本的类别标记不同的概率。$Gini$（S）越小，则数据集 S 的纯度越高。

2）样本集 S 被划分为 2 个子集 S_1 与 S_2，此次划分的 Gini 指标见下式。在节点分裂时，每个划分的 Gini 指标 $Gini_{split}$（S）越小，表示划分越合理，数据集的纯度越高。

$$Gini_{split}(S)=\frac{|S_1|}{|S|}Gini(S_1)+\frac{|S_2|}{|S|}Gini(S_2) \tag{8}$$

单棵决策树模型能够处理多种数据类型，且能够较好抑制噪声与处理缺失值，但是较为复杂的分类规则也使决策树模型易陷入过拟合的状态。所以在其之上改进，引入随机森林的思想。首先在原始训练数据样本集 S 中随机抽取 k 个单独样本进行训练，下次抽取之前，都将之前的样本放回到原始训练集 S 中。每次抽取的 k 个单独样本都将生成 1 个决策树，每棵决策树都具有相同的分布类型，统计所有决策树，按特定方法输出森林的结果。最简单且常用的方法为线性集成：当输入数据时，对所有决策树的结果投票，每棵决策树占 1 票，哪种类型被投票最多，即为样本最终的输出结果。

网格搜索法可以通过对 RF 模型中的各个参数进行穷举遍历，最终获取的超参数组合最优，来实现 RF 模型的优化。虽然 Grid search 遍历所有的组合，具有耗时较久的缺点，但和手动调参相比，它既具备优秀的速度，又能使误差最小，所以本文采用 Grid search 穷举遍历来优化 RF 模型。

2.3 ARIMA-RF 模型

组合模型通常有 2 种方法：串联法和并联法。一般都进行时间序列预测的模型，采用串联法组合模型。串联法先用甲模型得到预测残差值 Re，再通过残差值 Re 建立乙模型，得到残差修正值，最后两者取合得预测结果。并联法则需要确定各组合模型的权重值，确定权重的常见方法有等权重法、简单加权平均法（式(9)）、预测误差平方和倒数法（式(10)）和误差方差均方倒数法（式(11)）。

$$W_i=\frac{i}{\sum_{i=1}^{m}i},\ i=1,\ 2,\ \cdots,\ m \tag{9}$$

$$W_j=\frac{D_j^{-1}}{\sum_{j=1}^{J}D_j^{-1}},\ j=1,\ 2,\ \cdots,\ J \tag{10}$$

$$W_j=\frac{E_j^{-1/2}}{\sum_{j=1}^{J}E_j^{-1/2}},\ j=1,\ 2,\ \cdots J \tag{11}$$

式（10）中，D_j 为第 J 个模型的误差平方和；式（11）中，E_j 为第 J 个模型的预测误差方差。

本文使用并联组合法中的简单加权法建立 ARIMA-RF 组合模型，该组合模型流程如图 3 所示。

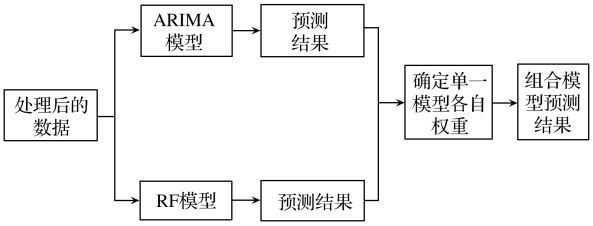

图 3 ARIMA-RF 组合模型流程

3 实例验证

3.1 ARIMA 实证分析

利用 2011 年 12 月 9 日至 2022 年 4 月 1 日的秦皇岛动力煤（Q5500）价格的周度数据建立时间序列图。图 4 为 2011 年 12 月 9 日至 2022 年 4 月 1 日的动力煤价格 X 的趋势图。由图可知，煤价序列无明显季节性波动，但无持续的增长或减少的趋势，应对数据进行平稳性分析。

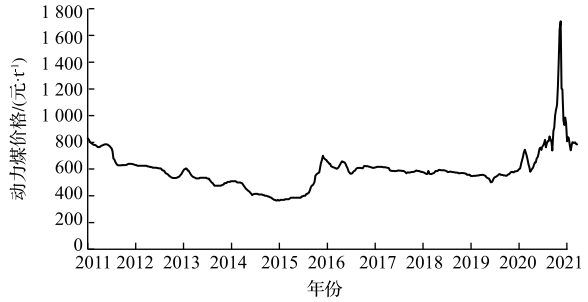

图 4 煤价趋势

通过单位根检验中的 ADF 方法验证序列是否具有平稳性。ADF 检验的结果为-2.25，5%的临界值为-2.87，结果明显大于 5%的值，显然不能显著地拒绝原假设，原假设为序列是非平稳序列；P 值为 0.19，且不十分接近于零。基于以上两者，可以认为原始序列是非平稳序列。

对原始非平稳序列 X 进行一阶差分，X'是差分后的序列，图 5 为序列 X'的趋势图，发现差分序列各值围绕均值上下波动，其 ADF 检验结果为-6.29，小于 1%的临界值-3.44，且 P 值为 3.67×10^{-8}，小于显著性水平（0.05），所以拒绝原假设。原假设为序列是不平稳序列，即一阶差分后序列平稳，因此该 ARIMA（p，d，q）模型的 d 取 1。

图 6 为序列 X'的自相关系数（ACF）和偏自相关系数（PACF）图。由图可知，ACF 和 PACF 图的截尾、拖尾都不显著。观察落于置信边界外的阶数，根据 PACF 图，可得 p 可选 0、1、2、3、4、6、8、9、10；根据 ACF 图，可得 q 可选 0、1、2、4、6、8、9。为给模型定阶，根据 BIC（贝叶斯信息）准则，通过网格搜索法，确定 BIC 最优的参数 p、q。BIC 准则公式如下。

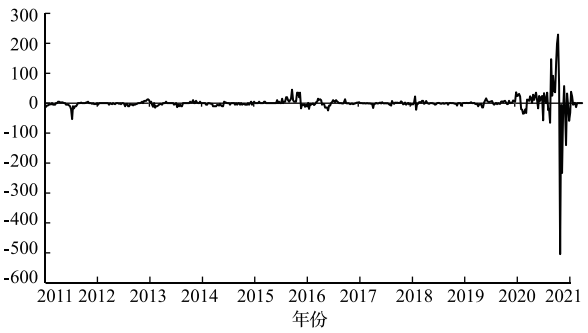

图 5 一阶差分序列 X'的趋势

$$BIC = k\ln(n) - 2\ln(L) \tag{12}$$

式（12）中的 k 是模型参数的个数，n 是样本数量，L 是似然函数。贝叶斯准则的结果越小越好，BIC 结果在 $p=2$ 且 $q=2$ 时最小，由此确定 $p=2$，$q=2$，p、q 的结果也在 ACF 和 PACF 图显示的范围中。所以得到 2011 年 12 月 9 日至 2022 年 4 月 1 日的动力煤价格时间序列模型 ARIMA（2,1,2）。

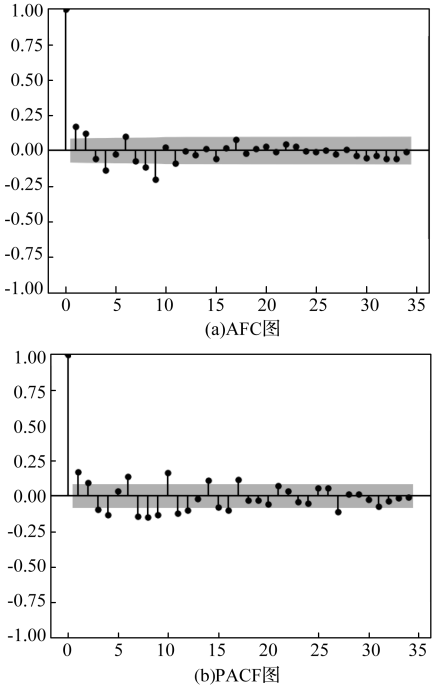

图 6 一阶差分序列 X'的 ACF、PACF 图

对 2011 年 12 月 9 日至 2022 年 4 月 1 日的动力煤价格的残差序列 X_t 绘制 ACF 图和 PACF 图，如

图 7 所示。图 8 则为残差序列 X_r 的 QQ 图。由图可知，序列 X_r 有短期相关性，相关系数趋向于零，残差是纯随机序列。

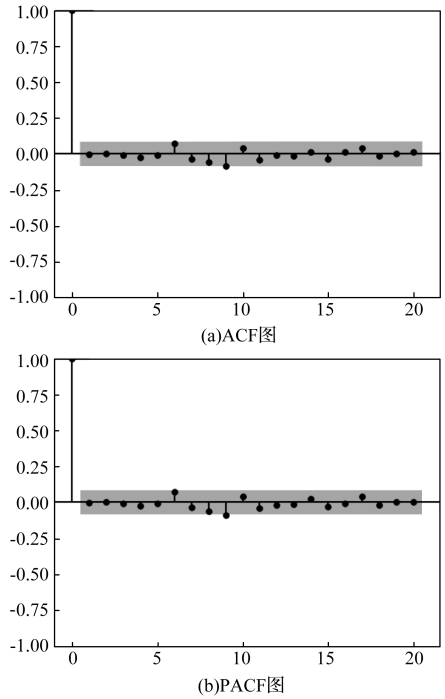

图 7 残差序列 X_r 的 ACF、PACF 图

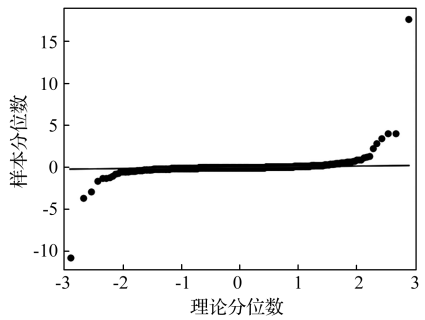

图 8 残差序列 X_r 的 QQ 图

对残差序列 X_r 进行 LB 检验（Ljung-Box test），P 值为 0.91，不小于显著性水平（0.05），所以接受原假设：相关系数为零，即相关系数和 0 没有显著性差异，残差序列 X_r 为纯随机序列。

用 ARIMA（2,1,2）模型预测 2022 年 1 月 7 日至 2022 年 3 月 25 日的秦皇岛动力煤（Q5500）价格，结果见表 6。煤炭价格的预测值与实际值的平均误差为 2.48%，是较为理想的结果。

2011 年 12 月 9 日至 2022 年 4 月 1 日的秦皇岛动力煤（Q5500）价格的 ARIMA（2,1,2）预测值与实际值的对比如图 9 所示。由图可知，10 年间秦皇岛动力煤（Q5500）价格在 ARIMA（2,1,2）模型下的预测值和实际值的趋势大体一致，但模型平均的均方误差（MSE）、均方根误差（RMSE）和平均绝对百分比误差（MAPE）分别为 939.05、30.64 和 0.91，结果不是很理想。综上，本文将引入 RF 算法，以寻求精度更高、兼顾线性和非线性关系的煤价预测模型。

表 6 ARIMA（2，1，2）模型预测结果和误差

时间	预测值	实际值	误差/%
2021-01-07	741	746.41	-0.73
2022-01-14	779	724.95	6.94
2022-01-21	802	798.66	0.42
2022-01-28	797	832.65	-4.47
2022-02-04	798	785.17	1.61
2022-02-11	799	767.03	4.00
2022-02-18	786	810.60	-3.13
2022-02-25	786	814.53	-3.63
2022-03-04	786	770.83	1.93
2022-03-11	786	758.54	3.49
2022-03-18	786	800.38	-1.83
2022-03-25	786	812.44	-3.36

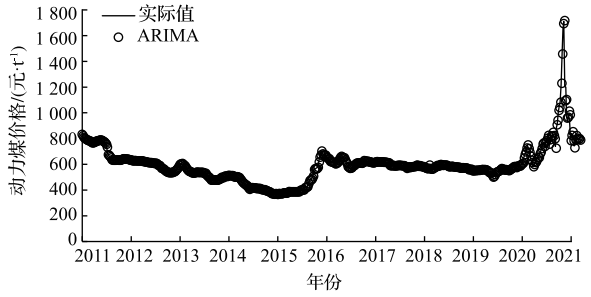

图 9 ARIMA（2，1，2）预测结果

3.2 ARIMA-RF 实证分析

采用公因子数据作为 RF 模型的输入，2011 年 12 月 9 日至 2022 年 2 月 25 日的动力煤价格为输出。遵循训练集与测试集之比为 3∶1 的原则，将 2011 年至 2019 年的动力煤价格作为训练集，2020 年至 2021 年的动力煤价格作为测试集。对建立的 RF 模型采用网格搜索法进行超参数优化，并检验泛化性能。RF 模型的 MSE、RMSE 和 MAPE 的值分别为 934.44、30.57 和 0.52。预测出下一周（2022 年 3 月 4 日）的煤价为 785.37，与真实值 786 相比，误差为-0.08%。使用 RF 模型预测 2021 年 4 月 2 日至 2022 年 2 月 25 日的动力煤价格，结果如图 10 所示。

基于以上优化后的 ARIMA（2,1,2）和 RF 模型，通过并联组合方法对两者的预测结果进行加权组合，得到组合后的 2011 年 12 月 9 日至 2022 年 2 月 25 日的动力煤价格预测结果。由于 2 个子模型预

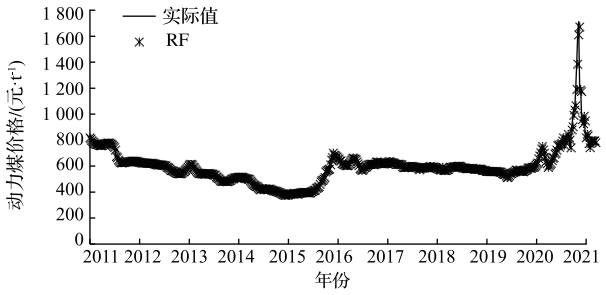

图 10 RF 预测结果

测的误差方差已知，不宜采用等权平均组合方法，故采用简单加权平均法更适宜。在简单加权平均法中，误差方差和权重系数成反比，误差小的预测值应具有更大的赋权比例，误差大的预测值将得到更小的赋权比例。本文通过 Python，以 MAPE 最小为目标，从 0~1 遍历，分别得到 ARIMA 模型的权重为 0.88，RF 模型的权重为 0.12。ARIMA-RF 模型的 MSE、RMSE 和 MAPE 分别为 924.94、30.41 和 0.17。采用 ARIMA-RF 模型预测 2021 年 4 月 2 日至 2022 年 2 月 25 日的动力煤价格，结果如图 11 所示。

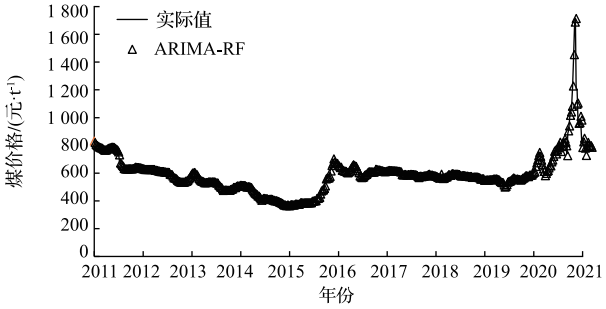

图 11 ARIMA-RF 预测结果

ARIMA-RF 模型预测的下一周（2022 年 3 月 4 日）的煤价为 785.63。ARIMA（2，1，2）、RF 和 ARIMA-RF 模型的预测结果如图 12 所示，3 个模型的评估指标见表 7。由图 12 可知，3 个模型 2012 年到 2022 年的拟合曲线与实际结果近似；由表 7 可知，组合模型的 MSE、RMSE、MAPE 值在 3 种模型中均为最小，因此该 ARIMA-RF 的组合模型用于动力煤价格的预测切实可行。

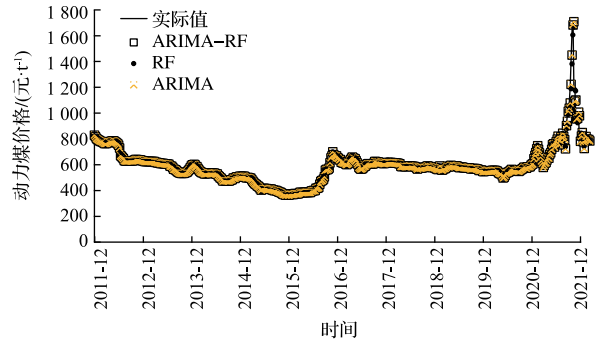

图 12 ARIMA（2,1,2）、RF 和 ARIMA-RF 模型的预测结果

表 7 3 个模型的评估指标

模型	MSE	RMSE	MAPE
ARIMA（2，1，2）	939.05	30.64	0.91
RF	934.44	30.57	0.52
ARIMA-RF	924.94	30.41	0.17

3.3 抽样对比分析

为验证本文组合模型具有更好的准确度，选择深度神经网络（DNN）模型、支持向量回归（SVR）模型、ARIMA 模型和 RF 模型作为参照模型，与 ARIMA-RF 模型作对比分析。DNN 神经网络模型和 SVR 模型作为已用于煤价预测的成熟机器学习模型，选择其作为对比组分析具有代表性；ARIMA 模型和 RF 模型作为参照组，描述了组合模型的适用性。由于 2021 年政策影响，限电停电导致 2021 下半年煤价变化幅度过大，所以选取 2021 年第三季度进行预测分析，同时分别抽选 2019 年、2020 年第 1、3 季度和 2021 年第 1 季度进行动力煤价格的滚动预测对比分析，如图 13 所示。

表 8 为 5 种模型 2019 年、2020 年和 2021 年第 1、3 季度的 RMSE。

表 8 动力煤价格预测模型的相对均方误差（RMSE）

模型	2019 年/季度		2020 年/季度		2021 年/季度	
	1	3	1	3	1	3
DNN	39.32	33.42	42.82	44.63	50.45	63.38
SVR	5.99	10.76	13.19	17.42	76.40	461.48
ARIMA	2.20	0.51	0.41	1.08	6.02	16.05
RF	3.65	4.93	3.60	5.95	6.88	22.74
ARIMA-RF	2.13	0.48	0.35	1.07	5.95	14.09

由图 13 和表 8 可知，ARIMA-RF 模型的准确性明显高于对照组的 4 种模型。通过图 13（a）与另外 5 个图的对比，可以看出在煤价较为平稳时，SVR 模型的预测效果优于 DNN 模型；而当煤价波动幅度较大时，SVR 模型的预测效果远低于 DNN 模型。ARIMA 模型和 RF 模型这 2 种模型的预测效果明显比 DNN 模型和 SVR 模型的预测效果更优秀，源于已经对前 2 种模型进行超参数优化，寻优后的 ARIMA 模型和 RF 模型拥有更高的准确性。基于 ARIMA 模型和 RF 模型组合而成的 ARIMA-RF 模型经过加权组合，在前两者基础上进一步降低了误差，在 5 种模型中具有最高的准确度。

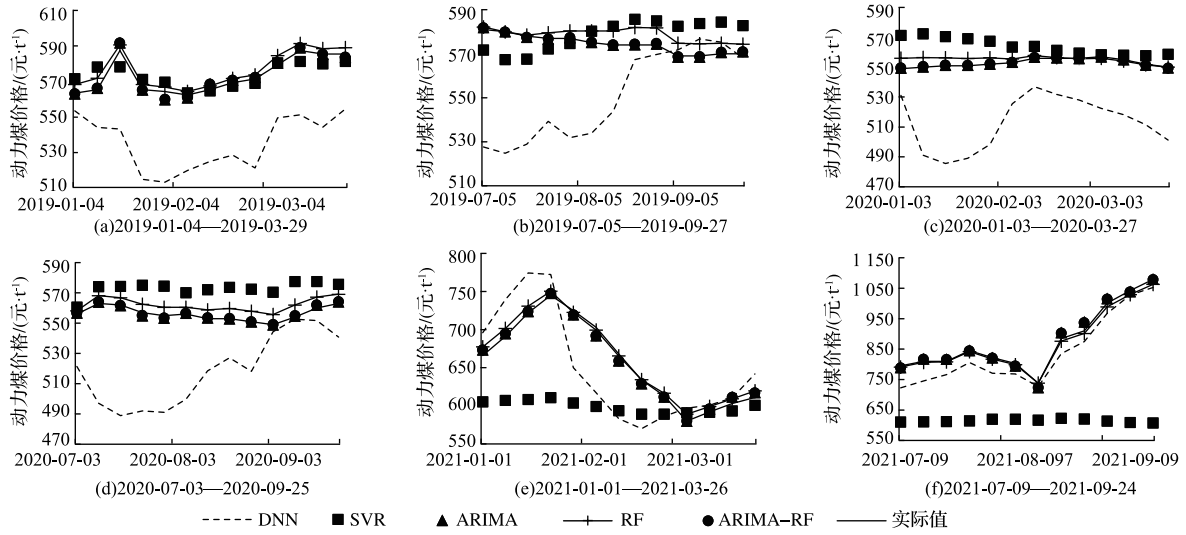

图 13 各模型的预测对比分析

4 结 论

针对国内动力煤价格非线性、受政策影响大和供需关系存在滞后性的问题，本文提出了一种基于 ARIMA-RF 的煤价预测组合模型，较经典的 DNN 模型、SVR 模型、ARIMA 模型和 RF 模型准确度更高，可为政府调控能源消费强度、深化能源体制机制改革、确保能源安全稳定供应和平稳过渡提供参考。通过本文研究，可以得到如下结论。

1）通过均值插补法和二次插值法进行数据清洗与转换，实现数据的预处理；根据皮尔森相关系数和类似主成分的因子分析，为动力煤价格的影响因素提供可靠性。

2）建立 ARIMA（2,1,2）模型和 RF 模型并优化，在此基础上，通过加权平均法得到子模型权重，建立 ARIMA-RF 组合模型。将其与子模型及经典 DNN 和 SVR 模型对比，验证出 ARIMA-RF 模型具有更好的回归预测结果及更高的准确度与精确度。

3）本文提出的 ARIMA-RF 模型可准确预测下一周煤价，缓解了滞后性对煤价的影响，可为企业合理预估煤价降低成本及政府制定政策提供一定参考，确保煤炭稳定供应，保障煤炭生产、流通、消费等环节平稳运行，有助于我国深化能源体制机制改革，早日实现“双碳”目标。

参考文献：

[1] Wen Lei, Diao Peixin. Simulation study on carbon emission of China's electricity supply and demand under the dual-carbon target[J]. Journal of Cleaner Production, 2022, 379: 134654.

[2] 舒印彪，赵 勇，赵 良，等. “双碳”目标下我国能源电力低碳转型路径[J]. 中国电机工程学报，2023，43（5）：1663-1672.

[3] 郭联哲，李晓军，谭忠富. 煤价波动对火电厂上网电价影响的数学模型及动态分析[J]. 电网技术，2005（7）：7-11.

[4] 廖志伟，黄杰栋，陈琳韬，等. 基于特征空间重构的差分-LSSVR 短期电煤价格预测[J]. 电网与清洁能源，2021，37（2）：1-10.

[5] 刘满芝，陈 梦. 基于 VAR-GARCH 模型的国内外煤炭价格动态互动关系研究[J]. 价格月刊，2017（3）：17-22.

[6] 许 晴，谭 鹏，张 成，等. 秦皇岛煤炭价格预测研究——基于因素分析法和支持向量机模型[J]. 价格理论与实践，2014（2）：79-81.

[7] 张克慧，牟博佼. 港口动力煤价格模型[J]. 工矿自动化，2013，39（7）：30-38.

[8] 刘亚成，马 磊. 发电公司燃煤价格预测模型[J]. 华东电力，2010，38（12）：1972-1974.

[9] 王帮俊，赵佳璐. 基于相空间重构的煤炭价格时间序列的混沌特征研究[J]. 工业技术经济，2018，37（7）：45-50.

[10] 向 超. 基于 ARIMA-SVR 组合模型的动力煤价格预测与实证研究[D]. 北京：对外经济贸易大学，2019.

[11] Pan Sitong. Coal price prediction based on LSTM[J]. Journal of Physics: Conference Series, 2021, 1802 (4): 042055.

[12] Xu Xiaojie, Zhang. Yun coking coal futures price index forecasting with the neural network[J]. Mineral Economics, 2023, 36 (2): 349-359.

[13] Matyjaszek M, Fernández P R, Krzemień A, et al. Forecasting coking coal prices by means of ARIMA models and neural networks, considering the transgenic time series theory[J]. Resources Policy, 2019, 61: 283-292.

[14] Alameer Z, Fathalla A, Li Kenli, et al. Multistep-ahead forecasting of coal prices using a hybrid deep learning model[J]. Resources Policy, 2020, 65 (C): 101588.

[15] Yang Shaomei, Chen Dongjiu, Li Shengli, et al. Carbon price forecasting based on modified ensemble empirical mode decomposition and long short-term memory optimized by improved

whale optimization algorithm [J]. Science of the Total Environment, 2020, 716 (C): 137117.

[16] Fu Xiangwan, Tang Mingzhu, Xu Dongqun, et al. Forecasting of steam coal price based on robust regularized kernel regression and empirical mode decomposition [J]. Frontiers in Energy Research, 2021, 9: 752593.

[17] Ho S L, Xie M, Goh T N. A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction[J]. Computers & Industrial Engineering, 2002, 42 (2): 371-375.

[18] Liu Xinrong. Research on the forecast of coal price based on LSTM with improved Adam optimizer[J]. Journal of Physics: Conference Series, 2021, 1941 (1): 012069.

[19] Ding Lili, Zhao Zhongchao, Han Meng. Probability density forecasts for steam coal prices in China: The role of high-frequency factors[J]. Energy, 2021, 220: 119758.

[20] Wang Xiaofei, Liu Chuangeng, Chen Shaojie, et al. Impact of coal sector's de-capacity policy on coal price [J]. Applied Energy, 2020, 265 (C): 114802.

[21] Zhu Shiqiu, Chi Yuanying, Gao Kaiye, et al. Analysis of influencing factors of thermal coal price[J]. Energies, 2022, 15 (15): 5652.

[22] Li Yanbin, Li Zhen. Forecasting of coal demand in china based on support vector machine optimized by the improved gravitational search algorithm[J]. Energies, 2019, 12 (12): 2249.

[23] Zhang Lihui, Ge Riletu, Chai Jianxue. Prediction of China's energy consumption based on robust principal component analysis and PSO-LSSVM optimized by the tabu search algorithm[J]. Energies, 2019, 12 (1): 196.

[24] 廖志伟，陈琳韬，黄杰栋，等. 基于特征空间变换与 LSTM 的中短期电煤价格预测[J]. 东北大学学报（自然科学版），2021，42（4）：483-493.

[25] Sohrabi, Parviz, Jodeiri Shokri, et al. Predicting coal price using time series methods and combination of radial basis function (RBF) neural network with time series[J]. Mineral Economics, 2021, 36 (2): 207-216.

[26] Feng Xiwen, Xin Mingshang. Analysis of factors affecting long-term coal demand in China Error correction model based on co-integration[J]. IOP Conference Series: Earth and Environmental Science, 2021, 769 (3): 032076.

[27] Gao Luhui, Wang Guoqing. Research on the improved neural network of coal price forecast based on co-integration theory[J]. IOP Conference Series: Earth and Environmental Science, 2021, 769 (4): 042028.

作者简介：季凌雲（1999—），女，江苏泰州人，硕士研究生，主要研究方向为热动力灾害预警、机器学习。E-mail: jilingyuncn@163.com

责任编辑：柳 妮