

基于DBSCAN-BIN算法的风功率曲线构建方法及应用研究

李进友*, 黄诗展, 黄金宇, 韦茂馨, 谢证泽, 韦伟昭

广西民族师范学院数学与计算机科学学院, 广西 崇左

收稿日期: 2026年4月25日; 录用日期: 2026年5月18日; 发布日期: 2026年5月25日

摘要

针对风功率曲线建模中异常数据干扰与传统方法平滑性不足的问题, 提出了一种基于DBSCAN-BIN的风功率曲线建模方法。首先, 为有效清洗原始数据中的异常数据, 提出基于分位数回归与 3σ 准则的异常数据处理方法, 进而提升数据分布的合理性与表征性。其次, 为精准剔除高风速段离散异常点, 基于DBSCAN密度聚类算法提出一种高风速段离散异常数据处理方法。再次, 利用BIN方法分箱统计构建风功率曲线, 实现降噪与拟合的协同优化。最后, 基于MAE、MAPE、RMSE及 R^2 模型评估指标, 与BIN、KNN、RF、DTR等方法进行实验对比。实验结果表明: 所提方法与传统方法相比评估指标更佳。其中, MAE为28.6177、MAPE为0.096、RMSE为39.4059、 R^2 为0.9934。所提方法构建曲线全局过渡无突变、局部无冗余波动, 对异常数据敏感度低, 为风功率预测、机组性能评估、故障诊断及运维调度提供可靠支撑。

关键词

风电机组, 风功率曲线, DBSCAN聚类, BIN方法, 曲线拟合

Research on Construction Method and Application of Wind Power Curves Based on the DBSCAN-BIN Algorithm

Jinyou Li*, Shizhan Huang, Jinyu Huang, Maoxin Wei, Zhengze Xie, Weizhao Wei

School of Mathematics and Computer Science, Guangxi Minzu Normal University, Chongzuo Guangxi

Received: April 25, 2026; accepted: May 18, 2026; published: May 25, 2026

*通讯作者。

文章引用: 李进友, 黄诗展, 黄金宇, 韦茂馨, 谢证泽, 韦伟昭. 基于 DBSCAN-BIN 算法的风功率曲线构建方法及应用研究[J]. 建模与仿真, 2026, 15(5): 106-120. DOI: 10.12677/mos.2026.155075

Abstract

To address the issues of abnormal data interference in wind power curve modelling and the insufficient smoothness of traditional methods, this paper proposes a DBSCAN-BIN based wind power curve modelling method. Firstly, to effectively remove abnormal points from raw data, an outlier processing method based on quantile regression and the 3σ criterion is developed, which improves the rationality and characterisation capability of data distribution. Secondly, to accurately eliminate discrete outliers in the high wind speed interval, a processing strategy for discrete abnormal data is presented based on the DBSCAN density clustering algorithm. Thirdly, the BIN method is employed to construct the wind power curve via binning statistics, realizing the coordinated optimisation of noise reduction and curve fitting. Finally, taking MAE, MAPE, RMSE and R2 as model evaluation metrics, comparative experiments are carried out with BIN, KNN, RF, DTR and other traditional algorithms. The experimental results demonstrate that the proposed method achieves superior overall performance over conventional methods. The corresponding evaluation indexes are MAE of 28.6177, MAPE of 0.096, RMSE of 39.4059 and R2 of 0.9934. The wind power curve established by the proposed method presents smooth global variation without abrupt changes and free of redundant local fluctuations, and shows low sensitivity to abnormal data. This method can provide a reliable reference for wind power prediction, wind turbine performance evaluation, fault diagnosis, as well as operation and maintenance scheduling.

Keywords

Wind Turbine, Wind Power Curve, DBSCAN Clustering, BIN Method, Curve Fitting

Copyright © 2026 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

目前,在我国 100 亿吨二氧化碳的排放中,因发电产生的二氧化碳约占 45% [1]。随着“双碳”目标的推进,新能源发电在电网中的占比不断提升[2]。风能作为一种可再生能源,在环境保护和可持续发展中具有很大的优势,而风电机组是将风能转化为电能的关键装备[3]。因此,高效地利用风能关键在于风电机组能够高效地将其转化为电能。而风功率曲线能够直观地反映风电机组在不同状态下的运行情况。基于风电机组的输出功率与环境风速的曲线模型对风电机组的性能与健康状态进行评估是一种常用方法[4]。针对功率曲线模型构建,风电领域专家、学者开展了如下研究:

在数据采集与预处理方面,文献[5]提出基于 MADM-QM 的风功率异常数据处理方法,实现横纵向异常数据的高效精准清洗。文献[6]基于堆积型异常的识别优势与四分位法对离散型异常的筛选能力,实现风电场复杂异常数据的精准识别。文献[7]利用 DBSCAN、四分位法识别堆积型及特征不明显的异常点,完整保留风功率曲线全程特性。文献[8]提出结合组合型异常识别与图注意力网络(GAT)生成对抗网络(GAN)的风电功率异常数据处理方法,利用 GAT 提取时空特征,结合梯度惩罚机制的改进型 GAN,实现异常数据精准重构。然而,上述文献在风电数据预处理上存在着对不同风速段异常点的剔除能力不足、难以平衡异常值清洗度与数据特征保留度的问题。因此,针对风功率数据中多类型异常值共存、高抗干扰性与高数据利用率难兼顾的问题,提出基于分位数回归、 3σ 准则与 DBSCAN 方法的数据处理方法。

在模型构建与拟合方面,文献[9]提出基于 XGBoost-Bin 的自动功率曲线极限算法。文献[10]通过偏

最小二乘法对得到机组多个变量与输出功率的相关程度,再以 PLS 构建功率曲线模型。文献[11]通过 K-means 聚类算法将不同机组划分,再提取各机组的实测数据,构建风电场输出功率模型。文献[12]提出了基于一种五参数和六参数逻辑函数的风功率速数据拟合方法。文献[13]针对解决单一变量模型无法适配复杂工况的缺陷,提出了基于高斯混合联结模型(GMCM)、回归人工神经网络(ANN)和贝叶斯人工神经网络(BANN)的功率曲线多元模型等。然而,上述文献在风功率曲线模型构建上存在着参数调试繁琐、模型复杂高导致工程实用性低的问题。因此,针对工程应用受限的问题,提出了基于 DBSCAN 聚类与 BIN 方法融合的风功率曲线模型构建方法。

在模型验证与评价方面,学者们普遍使用平均绝对误差(MAE)、均方根误差(RMSE)、平均绝对百分比误差(MAPE)、决定系数(R^2)对风功率曲线模型的性能进行全面、客观地定量评价。因此,将基于 MAE、MAPE、RMSE、 R^2 评估指标对所提方法构建的模型进行评价。

综上所述,为提升风功率曲线建模的精度、优化风电机组在高风速区间与低风速区间的曲线形态,提出一种基于 DBSCAN-BIN 算法的风功率曲线构建方法。首先,基于分段处理的方法通过分位数回归筛选出高置信区域数据,以 3σ 准则优化区域数据。其次,基于 DBSCAN 聚类算法进一步处理高、低风速区间存在的离散值,再基于 BIN 算法构建风功率曲线。最后,通过 MAE、MAPE、RMSE、 R^2 评估指标,对模型拟合度进行对比分析。

2. 风功率曲线数据处理

风电机组运行过程中受各种因素影响,SCADA 数据存在较多异常值。预处理完成后,首先使用分位数回归方法,基于风速与功率的关联特性,划定每个风段内高置信区间,优先剔除与风速完全不匹配的极端数据,为后续 3σ 处理划定合理的数据范围,避免分布偏差导致的 3σ 准则失效。经分位数回归粗过滤后,对已具备核心分布特征的数据执行 3σ 准则,可针对每个风段的功率分布特性,精准剔除风段内的边缘值,同时最大程度保留数据的整体分布特征避免被破坏。前两步处理已解决大部分堆积型、纵向异常带问题,但高风速段仍存在分散性、低密度的离散点,只有通过 DBSCAN 聚类的特性,通过整体聚类精准识别高风速段的离散点,同时完整保留剩余风速段的有效数据簇,弥补前两种统计方法的短板,最终为 BIN 方法构建平滑曲线提供高密度数据集。

因此,基本步骤为,首先对预处理后的数据进行分段操作。再根据高置信区间阈值 0.95 与 0.05 构建风速与功率的分位数回归曲线[14],选取两条分位数曲线之间的数据。再使用 3σ 准则保留 99.74% 以上的数据,进一步减少异常值并保留数据的主体特征,最后通过 DBSCAN 聚类进行深度剔除。

2.1. 数据预处理

由于发电机与齿轮箱工作时的发热、风向测量器的高度和距离与扇叶不同、特殊地形等因素,导致真实风速与测量风速之间存在误差。因此,为提高数据可利用率、降低数据误差,基于 IEC61400-1 标准[15]将机组“10 min”等级数据中的缺失和在数值上不合理的数据进行针对性过滤和剔除处理,并基于空气密度对风速、功率进行标准化[16],进而减少环境变量的影响,确保风功率曲线建模的准确性和可比性。基于空气密度对风速、功率,进行修正,如式(1)、(2)所示:

$$v_i = \bar{v}_i \cdot \left(\frac{\bar{\rho}_i}{\rho} \right)^{\frac{1}{3}} \quad (1)$$

$$p_i = \bar{p}_i \cdot \frac{\rho}{\bar{\rho}_i} \quad (2)$$

其中 $\bar{\rho}_i$ 为第条数据的空气密度均值, \bar{p}_i 为第 i 条数据的有功功率均值, \bar{v}_i 为第 i 条数据的风速均值, p_i

为第 i 条数据修正后的有功功率, v_i 为第 i 条数据修正后的风速, ρ 为标准大气压下的空气密度。

2.2. 数据处理

2.2.1. 基于分位数回归方法处理

为了精准捕捉正常数据的波动范围同时保留核心分布特征, 基于分位数回归方法进行处理, 具体步骤如下:

(1) 风功率数据分段上, 为了在局部风速区间内更精确地描述风速与功率的关系, 首先对原始风速数据进行等宽分段处理, 将全局回归问题转化为多个局部回归问题。设定间距为 $\omega = 0.5 \text{ m/s}$ [17], 并根据数据集的最大风速 v_{\max} 和最小风速 v_{\min} 自动生成连续的 m 个风速区间:

$$B_k = \begin{cases} [v_{\min} + \omega, v_{\min} + 2 \cdot \omega), k = 1 \\ [v_{\min} + 2\omega, v_{\min} + 3 \cdot \omega), k = 2 \\ \dots \\ [v_{\min} + m \cdot \omega, v_{\max}], k = m \end{cases} \quad (3)$$

(2) 构建分位数回归曲线, 在每个风段内进行处理。构建 $\tau = 0.05$ 和 $\tau = 0.95$ 即 5%分位数和 95%分位数的两个分位数回归模型。对第 k 个风段, 目标函数为:

$$\hat{\beta}_{k,\tau} = \arg \min_{\beta_{k,\tau} \in R^2} \sum_{i=1}^n L_{k,\tau} (p_i - (\beta_{k,\tau,0} + \beta_{k,\tau,1} v_i)) \quad (4)$$

其中 $(v_i, p_i) \in B_k$, $\hat{\beta}_{k,\tau} = [\beta_{k,\tau,0}, \beta_{k,\tau,1}]^T$ 为在该风段内在分位数 τ 水平下的最优回归参数, $\beta_{k,\tau,0}$ 为截距项, $\beta_{k,\tau,1}$ 是风速 v_i 的系数。 $L_{k,\tau,i}$ 为在该风段内第 i 条数据在分位数 τ 水平下的损失值。 n 为该风段内所有数据量。使用加权绝对偏差损失函数(CheckLoss):

$$L_{k,\tau,i}(u_i) = \begin{cases} \tau \cdot |u_i|, & u_i \geq 0 \\ (1-\tau) \cdot |u_i|, & u_i < 0 \end{cases} \quad (5)$$

其中 $u_i = p_i - \hat{p}_{i\tau}$ 为残差, p_i 为风电机组实测功率, $\hat{p}_{i\tau}$ 为 τ 分位数对应的功率预测值。基于损失函数训练, 最终得到该风段内的分位数回归曲线:

$$\hat{p}_{i\tau} = \beta_{k,\tau,0} + \beta_{k,\tau,1} \cdot v_i \quad (6)$$

其中, 当 $\tau = 0.05$ 时为下边界, 当 $\tau = 0.95$ 时为上边界。

(3) 取值, 构建出每个风段内的分位数回归上、下边界曲线, 并取落在两线之间的数据, 即: $(P', V') = \{(v_i, p_i) | p_{0.05i} \leq p_i \leq p_{0.95i}\}$ 。

2.2.2. 基于 3σ 准则处理

为了在深度去除异常值与最大程度保留原始数据分布特征之间实现平衡, 承接 2.2.1 分位数回归的等宽风速分段结果, 在每个风段内分别执行 3σ 准则。在每个风段内, 以风功率为对象, 计算该风段内风功率的均值 μ 与标准差 σ , 保留 $[\mu - 3\sigma, \mu + 3\sigma]$ 区间内的数据, 剔除区间的边缘值, 保留了该风段内 99.74% 的有效数据, 实现异常值清洗与数据特征保留的平衡。

3. 基于 DBSCAN-BIN 算法的风功率曲线构建方法

在国际功率特性评估标准 IEC-61400-12-1、IEC-61400-26-2 [12] 中使用的数据分箱法(Binning), 简称 BIN 方法, 对风功率曲线进行建模。但单纯的 BIN 方法建模可能存在模型曲线不光滑、数据易流失、对离散值较敏感等问题。因此, 提出一个基于 DBSCAN 聚类算法得出高密度风功率数据区域并通过 BIN 方法构建曲线的方法。避免传统 BIN 方法在数据稀疏区域的过拟合问题, 从而提高对数据利用的充分性,

进一步提高风功率曲线模型拟合的精确度。

3.1. DBSCAN 算法原理

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)是一种基于密度的聚类算法。在给定的数据集中，以每个数据点为中心，在二维数据中根据以 ϵ 为半径的圆形区域其他数据点的密度情况，将数据点分为核心点(Core Point)、边界点(Border Point)和噪声点(Noise Point)。通过调整邻域半径(Eps)和成为核心对象的在领域半径内的最少点数(Minpts)，能够有效识别任意形状的簇(Cluster)，并自动区分噪声点[18]。DBSCAN 算法处理基本步骤如图 1 所示：

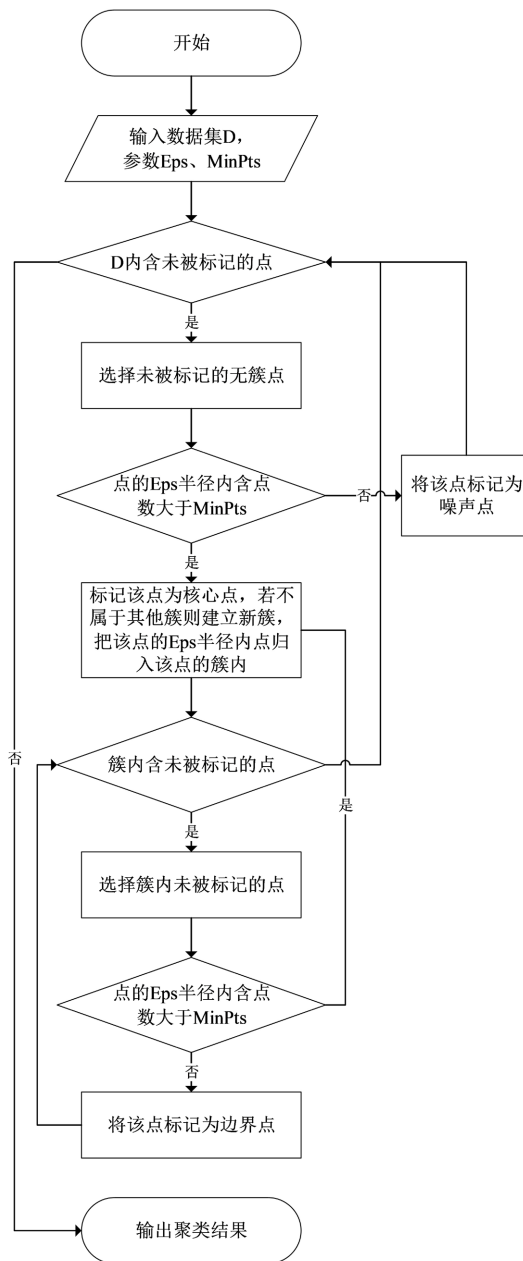


Figure 1. Flow chart of DBSCAN algorithm processing steps
图 1. DBSCAN 算法处理基本步骤图

3.1.1. 参数 MinPts 的选择

MinPts 的选择上, 该参数的核心作用是平衡聚类的抗噪性与对低密度簇的识别能力, 通用最优经验准则为取数据维度的 2 倍作为 MinPts 的取值[18], 即对二维数据集取 $\text{MinPts} = 4$ 。该取值既可避免 MinPts 过小导致的噪声点被误判为核心点与簇过度分裂问题, 也可防止 MinPts 过大导致的低密度有效簇被误判为噪声问题。

3.1.2. 参数 Eps 的选择

Eps 的选择上, 采用 k-距离图进行自适应确定, 该方法为 DBSCAN 参数选择的主流通用方法, 步骤与计算如下:

(1) 采用欧氏距离作为样本间的相似度度量, 对于二维空间中的任意两个样本 $x_i, x_j \in R^2$, 其距离计算公式为:

$$\text{dist}(x_i, x_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2} \quad (7)$$

其中 $x_{im}, (m=1,2)$ 为第 i 个样本第 m 维度的值。

(2) k-距离计算, 取 $k = \text{MinPts} - 1$, 即 $k = 3$ 。对数据集中的每个样本 x_i 计算其到第 k 近邻样本的距离, 记为 $\text{dist}_k(x_i)$, 称为该样本的 k-距离。

(3) 拐点识别, 将所有样本的 k-距离按升序排列, 以样本索引为横轴、k-距离为纵轴绘制 k-距离曲线。存在曲线从平缓到陡峭突变的拐点, 即为区分核心点与噪声点的最优分界点, 该拐点对应的 k-距离值即为 Eps 的最优理论取值。拐点采用 KneeLocator 算法实现自动识别, 灵敏度参数 α 用于平衡拐点识别的精准度与鲁棒性, 通用取值范围为 $\alpha \in [1, 2]$ 。

3.1.3. 参数敏感性分析

为验证所选参数的稳定性, 进行参数敏感性分析。固定 $\text{MinPts} = 4$ 为理论最优值, 在 k-距离图拐点处所取 Eps 值的合理波动区间内遍历取值, 从三个维度量化参数变动对聚类效果的影响, 各指标的通用计算公式如下:

(1) 整体轮廓系数(Silhouette Coefficient) S , 一种用于评估聚类算法效果的指标, 其值介于-1 到 1 之间。对于样本点 x_i 其轮廓系数 $s(x_i)$ 计算方法为:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}} \quad (8)$$

其中, $a(x_i)$ 是样本点 x_i 到其所属簇内其他所有样本点距离的平均值, 称为簇内不相似度。 $b(x_i)$ 是样本点 x_i 到其他各簇所有 n 个样本点距离的最小平均值, 称为簇间不相似度。整体轮廓系数 S 为所有样本点轮廓系数的平均值:

$$S = \frac{1}{n} \sum_{i=1}^n s(x_i) \quad (9)$$

(2) 簇数量 $N_{cluster}$, 作为聚类结果的直接体现, 反映了数据集中存在的自然分组数量。在 DBSCAN 算法中, 簇数量是通过算法自动确定的, 取决于参数 MinPts 和 Eps 的选择。

(3) 噪声比 η , 噪声比是指被算法识别为噪声点 N_{noise} 的样本数量占总样本数量 n 的比例。计算公式为:

$$\eta = \frac{N_{noise}}{n} \quad (10)$$

3.2. BIN 方法原理

为了有效平滑处理由湍流等因素引起的功率随机波动, 从而提炼出风速与功率之间的特征关系, 对

风速进行区间划分，并计算每个区间内的平均功率。具体实现步骤如下：

(1) 风速区间分箱，根据风电机组的运行参数，设定切入风速 V_m ，切出风速 V_M ，根据：

$$N = 4 \times (V_M - V_m) \tag{11}$$

$$V_{step} = \frac{V_M - V_m}{N} \tag{12}$$

计算出风速区间 (V_m, V_M) 上的分箱数量 N 与单个箱体区间的宽度 V_{step} 进行等距分箱。

(2) 构建滑动窗口，为了保证曲线的连续性和平滑性，采用宽度为 $2V_{step}$ 的滑动窗口覆盖整个风速区间 (V_m, V_M) ，对第 k 个窗口，其对应的风速范围为： $(V_m + (k-1)V_{step}, V_m + (k+1)V_{step})$ ， $(k=1, 2, 3, \dots, N-1)$ ，该窗口的中心位置为： $V_m + kV_{step}$ 。

(3) 数据聚合与统计，对于第 k 个滑动窗口，筛选出所有风速落在窗口内的所有数据点由：

$$\bar{v}_k = \frac{1}{n} \sum_{i=1}^n v_i \tag{13}$$

$$\bar{p}_k = \frac{1}{n} \sum_{i=1}^n p_i \tag{14}$$

其中 n 为该窗口的数据量，得出该窗口的平均风速 \bar{v}_k 与平均功率 \bar{p}_k ，作为该窗口的期望值。最终得到全部滑动窗口期望数据集： (\bar{v}_k, \bar{p}_k) 。

(4) 曲线构建，将所得到的数据集 (\bar{v}_k, \bar{p}_k) 通过线性插值，将这些离散点构建形成最终的、连续的风电机组风功率曲线。

3.3. 基于 DBSCAN-BIN 算法的风电机组功率曲线构建方法

基于 DBSCAN-BIN 算法的风电机组功率曲线构建方法的实施流程如图 2 所示：

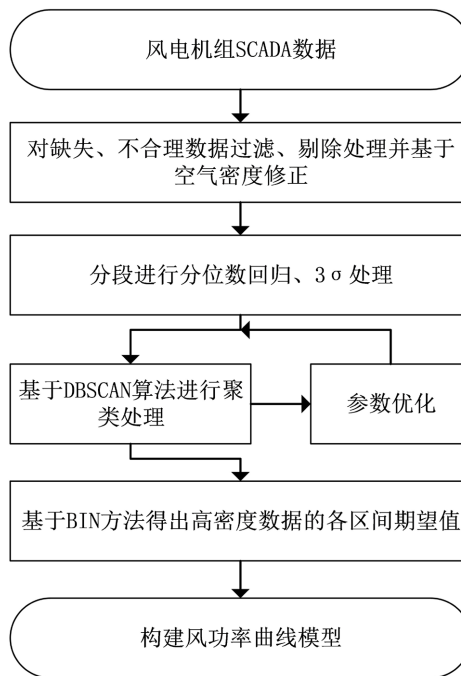


Figure 2. Construction of wind turbine power curve based on DBSCAN-BIN algorithm
图 2. 基于 DBSCAN-BIN 算法的风电机组功率曲线构建

如图 2 所示, DBSCANN-BIN 算法风电机组功率曲线模型构建步骤如下:

- (1) 对原始“10 min”等级数据中所含异常的数值进行处理, 并基于空气密度对风速和功率进行修正。
- (2) 采用分位数回归、 3σ 准则对预处理后的数据进行进一步处理。
- (3) 基于 DBSCAN 算法对数据进行聚类, 得到高密度的整体的数据集。
- (4) 基于 BIN 方法对高密度的数据集进行等宽分箱。根据小区间的风速均值和功率均值得出全部箱体的风速和功率期望值的数据集。
- (5) 基于箱体的风速和功率期望值的数据集, 构建风功率曲线模型。

3.4. 风电机组功率曲线模型性能评估

为更好评价曲线模型的准确性, 采用均方根误差(RMSE)、平均绝对误差(MAE)、平均绝对百分比误差(MAPE)和决定系数(R^2)等模型评估指标, 评价所提方法构建的风功率曲线模型性能与准确性, 同时对分析决策树回归模型(DTR)、随机森林(RF)、KNN 曲线构建模型准确性。评估指标计算方法如式(11)~(14)所示[19]:

$$MAE = \frac{1}{N} \sum_{i=1}^N |\hat{p}_i - p_i| \quad (15)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{p}_i - p_i)^2} \quad (16)$$

$$MAPE = \frac{1}{N} \sum_{i=1}^N \frac{|\hat{p}_i - p_i|}{p_i} \quad (17)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (\hat{p}_i - p_i)^2}{\sum_{i=1}^N (\hat{p}_i - \bar{p})^2} \quad (18)$$

其中, p_i 为测试集中第 i 条实测功率值, \hat{p}_i 为用 p_i 对应的风速 v_i 进行预测的模型预测功率值, \bar{p} 为测试集中实测功率均值, N 为测试集总数据量。

4. 实例分析

为检验所提 DBSCAN-BIN 算法构建风功率曲线模型的实用性与准确性, 选取某风电场额定 1.8 MW 型号风电机组($V_m = 2$ m/s, $V_M = 16$ m/s) 8 个月的 SCADA 运行数据进行实验研究, 结果如下。

4.1. 数据处理

基于 2.1 节预处理方法, 将实验原始训练数据中的机组状态位、有功均值、风速均值、空气密度均值中的空值、不合理值与突出异常值进行清洗。其次, 基于 2.2~2.3 节方法, 基于分段处理通过分位数回归、 3σ 准则把预处理后数据进一步处理, 实验结果如图 3 所示。原始数据中, 在功率为 0 时, 存在大量与之对应风速不匹配的异常值, 在风速中段(5~10 m/s)存在部分连续的纵向异常数据带, 通过预处理能够清洗功率为 0 时异常值。基于分位数回归与 3σ 方法对纵向异常数据带、高速段(10~20 m/s)数据带异常数据进行处理, 进而提高了风速段的数据密度。综上, 根据处理后的数据图像可直观地得出, 提出的数据处理方法既保证了风功率特征的准确性, 又实现了异常值的高效清洗。

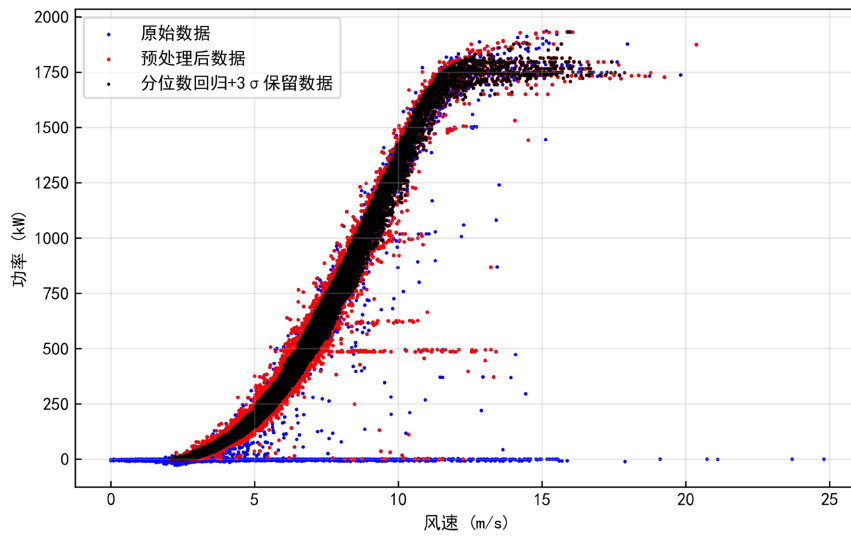


Figure 3. Scatter plot of processed data

图 3. 处理后数据散点图

4.2. DBSCAN 算法的参数选择与分析

基于 4.1 节最终处理后得到的数据，采用 Z-Score 标准化处理数据，并选取 $\text{MinPts} = 4$ ，进行以下选择与分析。

4.2.1. 参数 Eps 的选择

基于 3.1.2 的方法，得到 k-距离图并根据 KneeLocator 算法(取 $\alpha = 1.5$)找到拐点，得到在 $\text{MinPts} = 4$ 情况下，该数据集最佳 $\text{Eps} = 0.0373$ ，如图 4 所示。对于 KneeLocator 算法的灵敏度 α 的取值分析，绘制 α 对 Eps 值影响折线图，如图 5 所示。可见，当 α 取值在 $[1.1, 2.0]$ 时，曲线为稳定水平直线，故图 4 上所标拐点为可信拐点，此时 $\text{Eps} = 0.0373$ 的值为有效值。

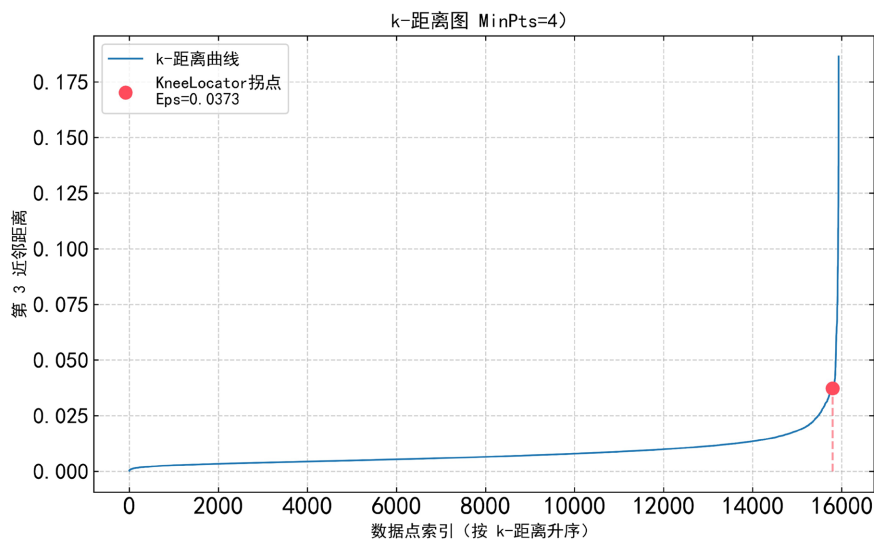


Figure 4. k-distance graph

图 4. k-距离图

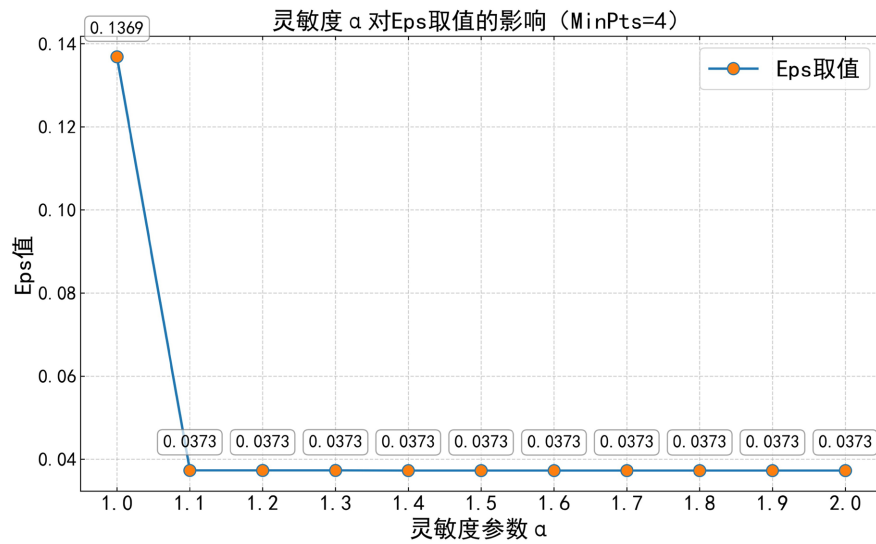


Figure 5. Effect of α on Eps
图 5. α 对 Eps 的影响图

4.2.2. 参数 Eps 的敏感性分析

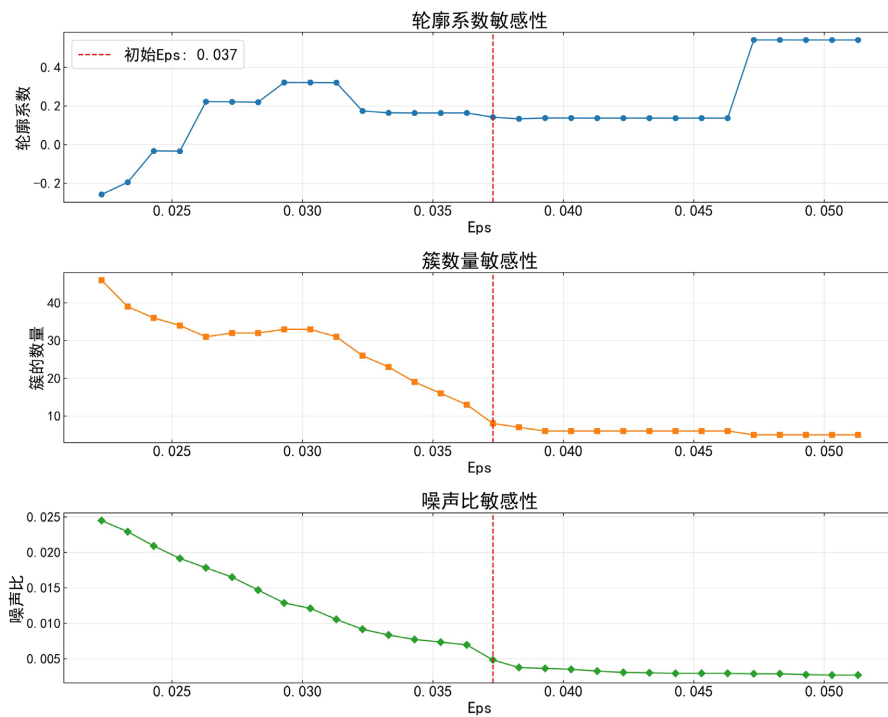


Figure 6. Eps sensitivity of evaluation indicators
图 6. Eps 敏感性指标图

基于 3.1.3 方法与 4.2.1 的结论, 在 $\text{MinPts} = 4$, $\text{Eps} = 0.037$ 情况下, 对 Eps 进行敏感性分析。绘制 $\text{Eps} = 0.037 \pm 0.015$ 区间内步长为 0.001 各值的各指标值折线图, 如图 6 所示。可观察到, 在轮廓系数敏感性上, 当 Eps 值较小时, 轮廓系数为负, 说明聚类效果较差。随着 Eps 值增加, 轮廓系数逐渐上升,

在 Eps 为 0.037 附近保持相对稳定的值。当 Eps 值继续增加到 0.05 左右时，轮廓系数急剧上升。在簇数量敏感性上，在 Eps 为 0.037 前，随着 Eps 值增加，簇数量明显逐渐减少。在 Eps 为 0.037 后，折线相对平缓，簇数量随 Eps 值变化的程度降低。在噪声比敏感性上，与簇数量敏感性的情况大致相同。

根据敏感性指标图，分析得到。在 Eps=0.037 附近，轮廓系数保持稳定，且处于簇数量和噪声比变化趋势改变的关键点。说明该值能够在聚类质量、簇数量和噪声比例之间取得较好的平衡。

4.3. 模型构建

基于 3.3 节曲线模型构建方法，基于 BIN 方法构建曲线。同时分别采用决策树回归(DTR)、随机森林(RF)、KNN 曲线进行曲线构建，实验结果如图 7~9 所示。

图 7 中，经过 DBSCAN 聚类后高风速段数据的密度增加，离散点基本被剔除。BIN 方法构建的曲线与所提方法构建的曲线在中低风速段基本重合，但在高风速段 BIN 方法构建的曲线受离散点影响产生较大波动，而所提方法构建的曲线更加平滑、连续程度更高。

图 8 图 9 为所提方法与其他方法构建曲线的对比图。其中，图 9(a)~(d)为所提方法分别与 BIN 方法、KNN 方法、RF 方法、DTR 方法构建的曲线的对比图。观察到：在图 8 中，整体上所提方法构建的曲线相比其他方法无突变，更加平滑、更符合数据整体分布特征。相比所提方法的拟合曲线，在图 9(a)中，传统 BIN 方法的拟合曲线在 >13 m/s 风速段出现连续的小幅波动。在图 9(b)中，KNN 方法的拟合曲线在 4 m/s~10 m/s 风速段内出现多处连续的震荡，在 >10 m/s 风速段出现极端波动。在图 9(c)~(d)中，RF 方法和 DTR 方法构建的曲线同时在 <10 m/s 风速段平滑度低，且在 >12 m/s 风速段几乎直接呈现水平直线，存在区间欠拟合。

综上，通过对比各方法曲线图，所提方法构建的曲线弥补了传统 BIN 方法对高风速段异常处理不足的短板，实现全风速区间的均衡拟合。同时，未出现 RF 和 DTR 方法构建模型的欠拟合情况，在全风速段始终保持平滑，无冗余波动。在含较多离散点的高风速段无 KNN 方法的极端突变，对离散值敏感度低，抗干扰能力较强。因此，从风功率曲线图看，所提方法优于传统 BIN、RF、DTR 及 KNN 方法，更贴合风电机组实际运行特性。

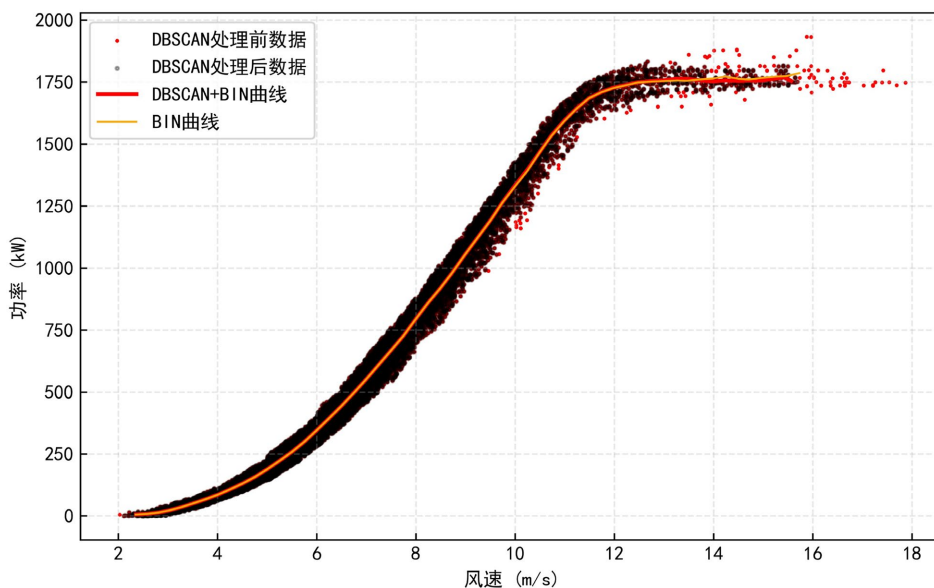


Figure 7. Comparison plot of curves constructed by BIN method and DBSCAN-BIN method
图 7. BIN 方法与 DBSCAN-BIN 方法构建曲线对比图

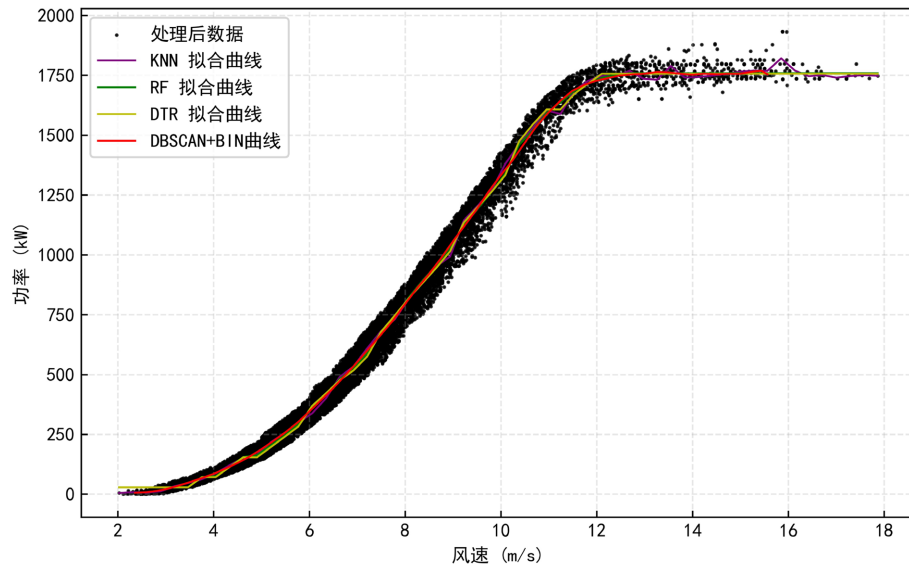


Figure 8. Comprehensive comparison plot of curves constructed by various models
图 8. 各模型构建曲线综合对比图

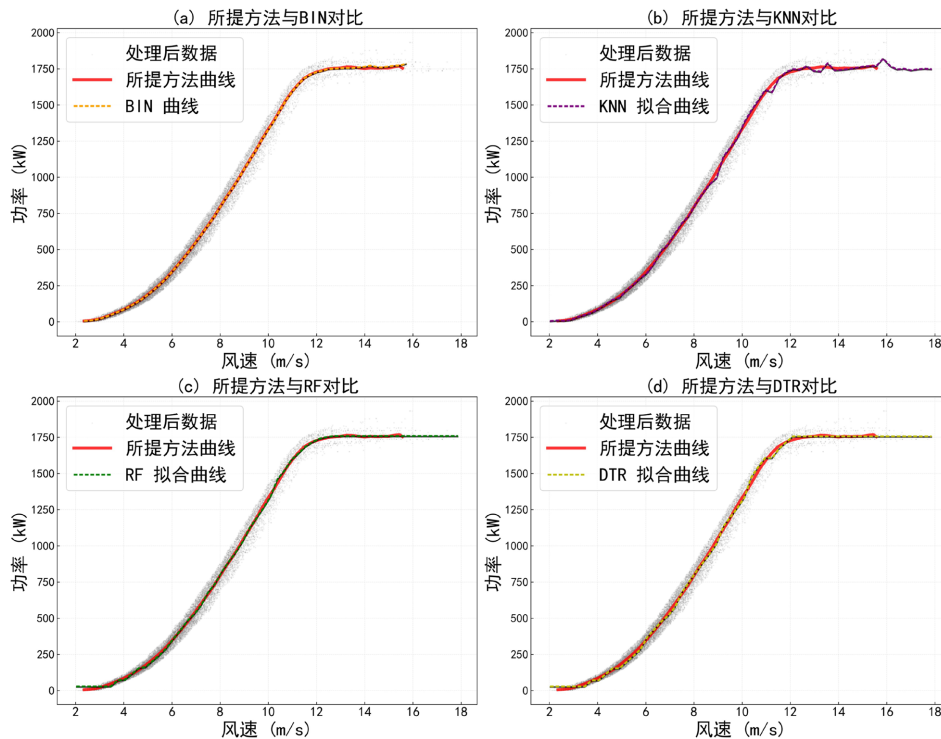


Figure 9. Comparison plot of curves constructed by various models and the proposed method
图 9. 各模型与所提方法构建曲线模型对比图

4.4. 模型评估指标

为客观、全面地量化各建模方法的拟合性能与误差水平，以 MAE、MAPE、RMSE 及 R^2 评估指标量化模型误差，进而验证所提 DBSCAN-BIN 方法在拟合精度与稳定性上的显著优越性，对比结果如表 1、

图 10 所示。

如表 1、图 10，所提方法略优于 BIN 方法，显著低于 KNN、RF 及 DTR，体现了在平均误差幅值控制上的优势。MAPE 中，所提方法与 BIN 方法持平，远优于 RF，说明其相对偏差水平更优。RMSE 的对比中，所提方法略低于 BIN 方法，较 KNN、RF 及 DTR 降低幅度明显，反映出误差离散程度的有效抑制。R² 方面，所提方法与 BIN 方法一致，高于 KNN、RF 及 DTR，表明其对数据整体分布特征的拟合一致性更优。

综上，通过 RMSE、MAE、MAPE 及 R² 四大核心指标的多维度量化评价，所提方法在误差控制与数据拟合优度上展现出显著优越性，其性能在与 BIN、KNN、RF 及 DTR 等主流方法的对比中，实现了绝对误差、相对误差、误差离散程度与拟合一致性的全面均衡，量化结果与前文定性分析的曲线平滑性结论形成了完整的论证闭环。

Table 1. Evaluation indicators of each model

表 1. 各模型评估指标

模型	指标	MAE	MAPE	RMSE	R ²
本研究		28.6177	0.0960	39.4059	0.9934
BIN		28.6473	0.0960	39.4148	0.9934
KNN		29.2339	0.0979	40.8535	0.9925
随机森林(RF)		29.8000	0.1445	40.2600	0.9932
决策树模型(DTR)		29.0700	0.0966	40.7000	0.9930

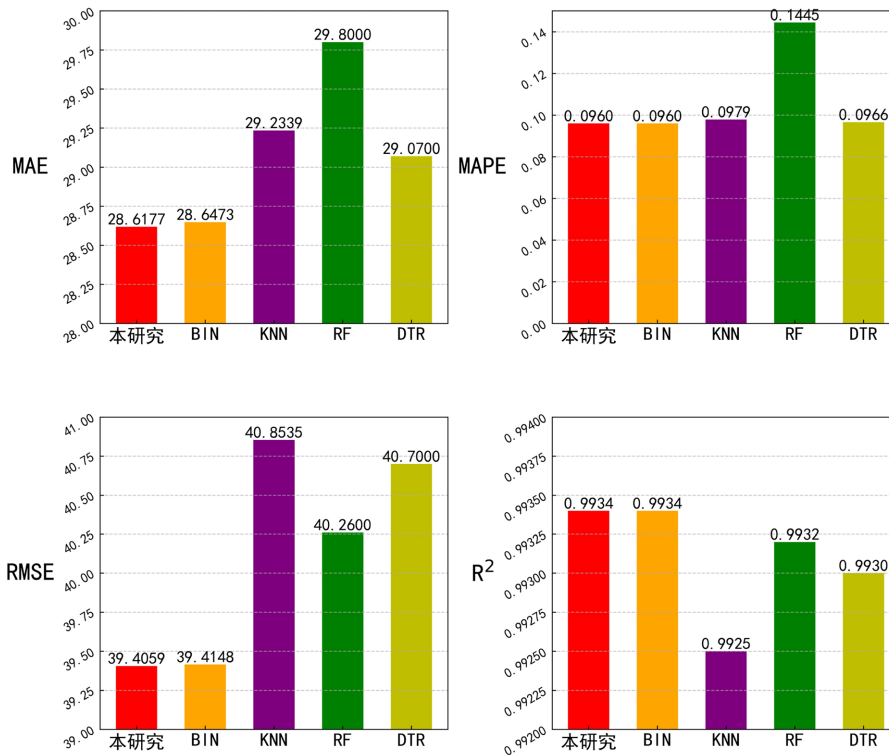


Figure 10. Comparison histogram of evaluation indicators for each curve model

图 10. 各曲线模型评估指标对比直方图

4.5. 模型应用分析

所提风功率曲线模型构建方法嵌入风电场 SCADA 系统的实时数据处理模块,可基于当前风速、空气密度等运行参数,快速输出精准的功率预测值,辅助运行人员提前调整机组运行策略,减少功率波动带来的损耗。通过模型输出的功率曲线与机组实际运行功率的偏差分析,结合 MAE、 R^2 等指标的长期变化趋势,可识别机组是否出现叶片磨损、传动系统老化等隐性故障,为预防性维护提供数据支撑。针对风电场内不同机组的功率曲线模型结果,可量化各机组的发电潜力差异,指导风电场进行机组运行优先级调度,同时为风电场的扩容或机组改造提供决策依据。

5. 结语

为解决风功率曲线建模中异常值干扰及传统方法平滑性不足的问题,提出了一种 DBSCAN-BIN 建模方法。在数据处理上,采用分段处理的方法以分位数回归、 3σ 准则处理数据,并利用 DBSCAN 聚类算法剔除不同风速段离散点,再以 BIN 方法构建曲线模型。通过图像和评估指标的对比分析,所提方法较 BIN、KNN、RF、DTR 传统方法在图像上更加平滑,全段无突变和冗余波动。在评估指标上,较最优传统方法的 MAPE 和 R^2 持平,而 MAE、RMSE 分别降低了 0.0296、0.0089。因此,所提方法为风功率曲线精准建模提供了兼具抗干扰性、稳定性与工程实用性的技术方案,可有效支撑风功率预测、机组性能评价及故障诊断等后续工程应用。

基金项目

广西民族师范学院科研项目(2024YB123); 广西大学生创新创业训练计划项目(S202510604090)资助项目。

参考文献

- [1] 王丽娟, 张剑, 王雪松, 等. 中国电力行业二氧化碳排放达峰路径研究[J]. 环境科学研究, 2022, 35(2): 329-338.
- [2] 刘明宝, 廖灿坤, 刘泽地. 南方区域新能源场站惯量响应研究与应用展望[J]. 红水河, 2025, 44(2): 70-76, 81.
- [3] Dai, J., Liu, D., Wen, L. and Long, X. (2016) Research on Power Coefficient of Wind Turbines Based on SCADA Data. *Renewable Energy*, **86**, 206-215. <https://doi.org/10.1016/j.renene.2015.08.023>
- [4] 张鑫淼. 基于 SCADA 数据的风电机组性能分析及健康状态评估[D]: [硕士学位论文]. 北京: 华北电力大学(北京), 2017.
- [5] 莫丰源, 王卫华, 郭前. 基于 MADM-QM 的风电机组风功率异常数据处理方法[J]. 可再生能源, 2025, 43(3): 339-345.
- [6] 吴永斌, 张建忠, 邓富金, 等. 基于方差变化率判据-四分位的风电场功率异常数据识别[J]. 电力工程技术, 2023, 42(4): 141-148.
- [7] 李宣谕. 新能源场站风功率曲线异常数据处理算法[J]. 安徽电气工程职业技术学院学报, 2023, 28(04): 62-72.
- [8] 钟吴君, 李培强, 涂春鸣, 等. 基于组合型异常识别与 GAT-GAN 的风电功率异常数据处理[J/OL]. 电力系统自动化: 1-21. <https://link.cnki.net/urlid/32.1180.TP.20250820.1451.009>, 2025-11-23.
- [9] 李进友, 李媛, 王海鑫, 等. 基于 XGBoost-Bin 自动功率极限计算的风电机组健康性能评估及预测[J]. 计算机集成制造系统, 2024, 30(6): 2172-2185
- [10] 解加盈, 郭鹏. 基于多变量选择的深度神经网络功率曲线建模[J]. 华电技术, 2019, 41(8): 27-31, 52.
- [11] 沈清野, 沈清坤, 秦本双. 基于聚类-实测风功率曲线的风电场等值模型研究[J]. 黑龙江电力, 2016, 38(2): 114-118.
- [12] Aldaoudeyeh, A., Alzaareer, K., Harasis, S., Al-Odat, Z., Obeidat, M., Mansour, A., *et al.* (2021) A New Method to Fit Logistic Functions with Wind Turbines Power Curves Using Manufacturer Datasheets. *IET Renewable Power Generation*, **16**, 287-299. <https://doi.org/10.1049/rpg2.12309>
- [13] Lázaro, R., Yürüşen, N.Y. and Melero, J.J. (2022) Wind Turbine Power Curve Modelling Using Gaussian Mixture

- Copula, ANN Regressive and Bann. *Journal of Physics: Conference Series*, **2265**, Article ID: 032083. <https://doi.org/10.1088/1742-6596/2265/3/032083>
- [14] Hyndman, R.J. and Fan, Y. (1996) Sample Quantiles in Statistical Packages. *The American Statistician*, **50**, 361-365. <https://doi.org/10.1080/00031305.1996.10473566>
- [15] IEC (2019) Wind Energy Generation Systems—Part 1: Design Requirements, Edition 4.0: IEC61400-1-2019. International Standard Norme Internationale, 1-172.
- [16] 任阔, 李洪奎. 两种修正风力发电机功率曲线的方法研究[J]. 输配电工程与技术, 2018, 7(4): 71-76.
- [17] 李闯, 田春华, 刘家扬, 等. 风力发电机组对风偏差检测算法研究与应用[J]. 可再生能源, 2020, 38(5): 620-624.
- [18] Ester, M., Kriegel, H.P., Sander, J. and Xu, X.W. (2009) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, 2-4 August 1996, 226-231.
- [19] Zhao, H., Wu, Q., Hu, S., Xu, H. and Rasmussen, C.N. (2015) Review of Energy Storage System for Wind Power Integration Support. *Applied Energy*, **137**, 545-553. <https://doi.org/10.1016/j.apenergy.2014.04.103>