**Computer Science and Application** 计算机科学与应用, 2021, 11(10), 2517-2528
Published Online October 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.1110255

Hans汉斯

# 基于DBSCAN的风机功率异常数据清洗

**孔维胜[1,2]，朱海鹏[1]，王晓东[1]，石明全[1,2*]**

[1]中国科学院重庆绿色智能技术研究院，重庆
[2]中国科学院大学，北京

收稿日期：2021年8月15日；录用日期：2021年10月12日；发布日期：2021年10月19日

## 摘　要

**风功率曲线是评价风电机组性能的重要指标，对风电场整体的运行管理具有重要意义。在风电机组实际运行过程中，由于设备故障及自然因素等原因的影响会导致数据采集与监视控制系统(SCADA)采集的数据中存在大量异常数据，导致风功率曲线评价不准确。本文从异常数据的产生机理分析，将数据分为0功率堆积数据、恒功率限电数据和分散型异常数据，根据不同类型数据特征，提出了基于DBSCAN和区间DBSCAN (DBSCAN-Interval DBSCAN)组合的异常检测模型，实现了对运行数据的清洗。最后，将本方法应用到某风场全年的风机采集数据中，对其进行数据清洗，结果表明该方法可以有效地检测和分离运行数据中的异常数据，在保证数据完整性的基础上提高了数据质量，显著提高了风电机组性能分析的准确性。**

## 关键词

**风电机组，异常检测，风功率曲线，密度聚类**

# Cleaning of Abnormal Data of Wind Turbine Power Based on DBSCAN

**Weisheng Kong[1,2], Haipeng Zhu[1], Xiaodong Wang[1], Mingquan Shi[1,2*]**

[1]Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing
[2]University of Chinese Academy of Sciences, Beijing

Received: Aug. 15th, 2021; accepted: Oct. 12th, 2021; published: Oct. 19th, 2021

## Abstract

**The wind power curve is an important indicator for evaluating the performance of wind turbines,**

*通讯作者。

文章引用: 孔维胜, 朱海鹏, 王晓东, 石明全. 基于 DBSCAN 的风机功率异常数据清洗[J]. 计算机科学与应用, 2021, 11(10): 2517-2528. DOI: 10.12677/csa.2021.1110255

**and is of great significance to the overall operation and management of the wind farm. However, due to equipment failures and natural factors, there will be a large number of abnormal data in the data collected by the data collection and monitoring control system (SCADA) in practices, making the wind power curve inaccurate. In this paper, aimed to clean the data of wind power curve, the data is divided into zero power accumulation data, constant power limit data and scattered abnormal data based on the analysis of the generation mechanism of abnormal data firstly. Then, according to the characteristics of the different types of data, a combination of DBSCAN and interval DBSCAN (DBSCAN-Interval DBSCAN) method is proposed, which realizes the anomaly operating data detection and cleaning. Finally, this method is applied to the annual wind turbine collection data of a wind farm to clean the data. The results show that DBSCAN-Interval DBSCAN method can effectively detect and clean abnormal data in the operating data, enforce the data integrity, and improve data quality, which significantly enhanced the accuracy of performance analysis of wind turbines.**

## Keywords

**Wind Turbine, Anomaly Detection, Wind Power Curve, Density Clustering**



## 1. 引言

碳达峰、碳中和(“双碳”)作为我国提出的两阶段碳减排目标，具有重要的战略意义。实现双碳目标是我国实现可持续发展的内在要求和构建人类命运共同体的责任担当[1]。

加强技术创新，推动新能源战略发展是实现双碳目标的重要途径。

风力发电是新能源战略发展的重点之一。在风力发电过程中，风电机组的运行数据会对风电场协调优化以及功率预测产生直接的影响，但是由于设备自身状况或者极端的自然因素等原因，风电机组运行数据的采集工作会受到影响。因此研究风电机组异常数据检测与清洗方法，对提升风电场运行维护的准确性和经济性有重要意义[2]，对双碳目标的实现有重要的推动作用。

当前在风机监测方法中，异常数据的检测与清洗是当前研究的热点，一些高校和机构对其进行了深入研究。朱倩雯等[3]采用四分位法进行处理，但在异常数据占比变大时，识别效果会变差；Wang [4]提出了基于的数据检测方法，并假设概率密度函数符合正态分布，但是这与实际风功率的概率密度函数状态不符，导致其通用性不高；娄建楼等[5]提出组内最优方差算法，虽然对堆积型数据有一定效果，但是个例性较强，应用的鲁棒性较差；沈小军[6]采用变点分组与四分位相结合的方法进行异常数据检测，但是存在正常数据误删的现象；Zhang [7]提出利用局部离群因子(Local Outlier Factor, LOF)算法，基于密度去筛选异常值点，可以很好地识别分散型异常数据，但是对堆积型数据不能有效识别。胡阳等[8]提出使用置信区间等效边界法进行异常数据的识别与清洗，但是边界的偏差会导致数据出现误删或者漏删的情况，对结果造成影响。

虽然当前风机数据清洗取得了显著的成果，但仍存在数据误删、通用性较差、鲁棒性差等问题。本文首先基于异常产生的机理分析了风电机组不同异常值的分布特征，并进行了分类。然后依据异常值分布特征提出了基于密度聚类与区间密度聚类(DBSCAN-Interval DBSCAN)组合的异常检测模型，最后从方法的有效性、合理性以及结果上进行了分析验证。

## 2. 异常数据特征

在风电场基于 SCADA 采集回的风电机组原始运行数据中，往往包含大量的异常数据。造成这些异常数据的本质是机组出现故障以及由于失修、弃风限电等多种原因出现的停机或者低性能运行导致；或者受到电磁干扰、风机脱网、极端天气等。由于不同原因造成的异常数据在风速－功率曲线上也会有不同的呈现。大致可分为底部 0 功率堆积数据、恒功率异常数据和周围分散性异常数据。各类数据大致分布情况如图 1 所示。

分析图 1 可知：

1) 曲线底部 0 功率堆积异常数据。此类数据通常由于机组故障、检修或者通信故障等原因造成。在该类情况下，机组的理论输出功率应该为零，但由于若风机的叶片未转动，但风机的测控系统需要电力驱动[9]，则导致数据中会出现为负值的情况。

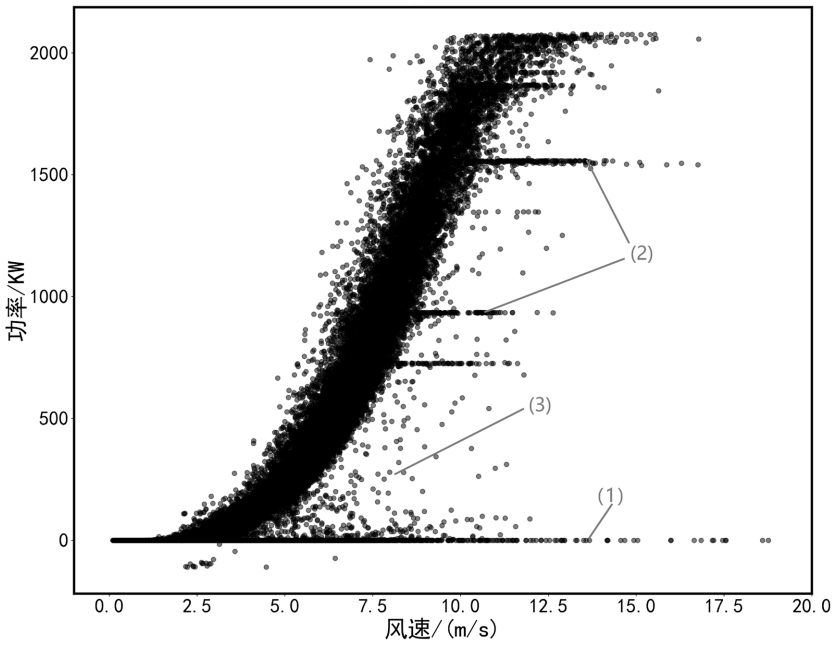



**Figure 1.** Diagram of distribution of Wind speed-Power abnormal data
**图 1.** 风速－功率异常数据分布示意图

2) 曲线中部恒功率堆积数据。这类数据在风速－功率曲线中常表现为一条或多条低于正常功率曲线的密集带。造成此类故障的原因常是弃风限电，可能因为按照调度计划要求，使机组一直按低于正常功率的状态运行，进而产生了恒功率堆积数据。在风电场实际运行过程中，由于当前电网消纳能力不足等原因，使得机组强制弃风已是常态[10]。

3) 周围分散型异常数据。这类异常数据在风速－功率曲线中变现为在功率曲线附近低密度且无规律的散点。此类异常数据的产生常由极端的天气、信号传播中的噪声等随机因素造成的，也正是因为随机因素的不确定性导致此类数据在功率曲线外随机分布。

根据上述几类异常数据的分布特征和产生原因可知，堆积型数据往往是由于设备故障和限风弃电等原因造成的，并且没有办法在短时间内恢复。而分散性异常数据则是由短时间可以恢复的状态或者故障产生的，这些因素是随机产生是，存在不确定性因此表现为分散状态。

针对分散型、堆积型异常数据的检测方法，本文将依次进行其原理解析。

## 3. 基于改进密度聚类的异常检测方法

### 3.1. DBSCAN 密度聚类方法

DBSCAN 密度算法是假设能够通过样本分布的紧密程度确定聚类结构[11]，并可以将一定密度的部分划分成簇。该算法可以由给定的参数 Eps 和 Minpts 来确定一个簇，其中 Eps 表示确定的邻域半径，Minpts 表示在以 Eps 确定的邻域中，核心点至少应该包含对象的数量。

在确定参数的前提下，DBSCAN 算法的思想如下：

1) 初始化样本集 D 中所有对象的状态，设为未访问；

2) 遍历样本集 D 的每个对象 t 进行判断，如果 t 已经被归入某个簇或被标记为异常值点，则结束判断，否则，继续执行；

3) 检测对象 t 的 Eps 邻域，如果包含的对象数不小于 Minpts，则标记对象 t 为核心点，并将本身以及邻域内所有对象放入新簇 C；

4) 对于 t 的 Eps 邻域内未被处理的对象，检查其邻域，并与 Minpts 比较，将不小于 Minpts 的对象，将其与其邻域中未归入簇的对象加入 C。

### 3.2. Interval DBSCAN 密度聚类方法

DBSCAN 方法对于边界不清晰的异常数据处理有一定的局限性，本文基于区间的密度聚类方法(Interval DBSCAN)，是将普通的密度聚类方法在全局参数上的聚类，划分在局部参数进行聚类。通过对数据某一变量按照一定范围进行拆分，将整体数据分成多组数据，再每组进行聚类，最终将每组结果进行整理，达到分离异常数据的目的。

### 3.3. DBSCAN-Interval DBSCAN 算法实现过程

针对风机 SCADA 数据分布的特点，提出利用二次密度聚类的方法将风机的运行数据进行清洗。具体流程如图 2 所示。

1) 首先将数据进行第一次密度聚类，此次聚类会识别出分散型的异常数据，但是由于底层堆积数据以及恒功率数据也具有高密度的特点，且与正常数据值界限不清晰，所有会存在未被识别的情况，将进入第二次密度聚类识别；

2) 第二次密度聚类方法利用分区域的方法，将全局数据进行拆分，在每个小区间利用 DBSCAN 算法。将数据集按风速进行划分，以风速 a 的间隔分成多个区域。得到每个小区间的数据集$U_i$为：

$$U_i = \left\{(v,p)\middle|(v,p)\in U, 0.5i < v < 0.5+0.5i\right\}, i\in(0,1,\cdots,n) \tag{1}$$

式中：$U_i$为第$i$个区间的数据集；$(v,p)$为落在第$i$个区间的元素。

3) 在每个小区间内使用密度聚类方法，对数据进行识别，可以有效的识别出未识别出的堆积数据。

## 4. 算例试验

### 4.1. 数据样本

为验证提出的方法的实用性，选取国内某风场一年的 SCADA 真实运行数据进行验证，该风场风机的基本参数为：额定功率 2000 kW，切入风速 3 m/s，切出风速 25 m/s。这里分别选取异常数据分布比较典型的 4#、5#号机组数据来进行验证。其连续一年的原始数据如图 3 所示。

4#机组数据，异常值多为周围分散型数据，以及 0 功率堆积异常数据；5#机组存在更多的限电数据，堆积异常数据特征明显。

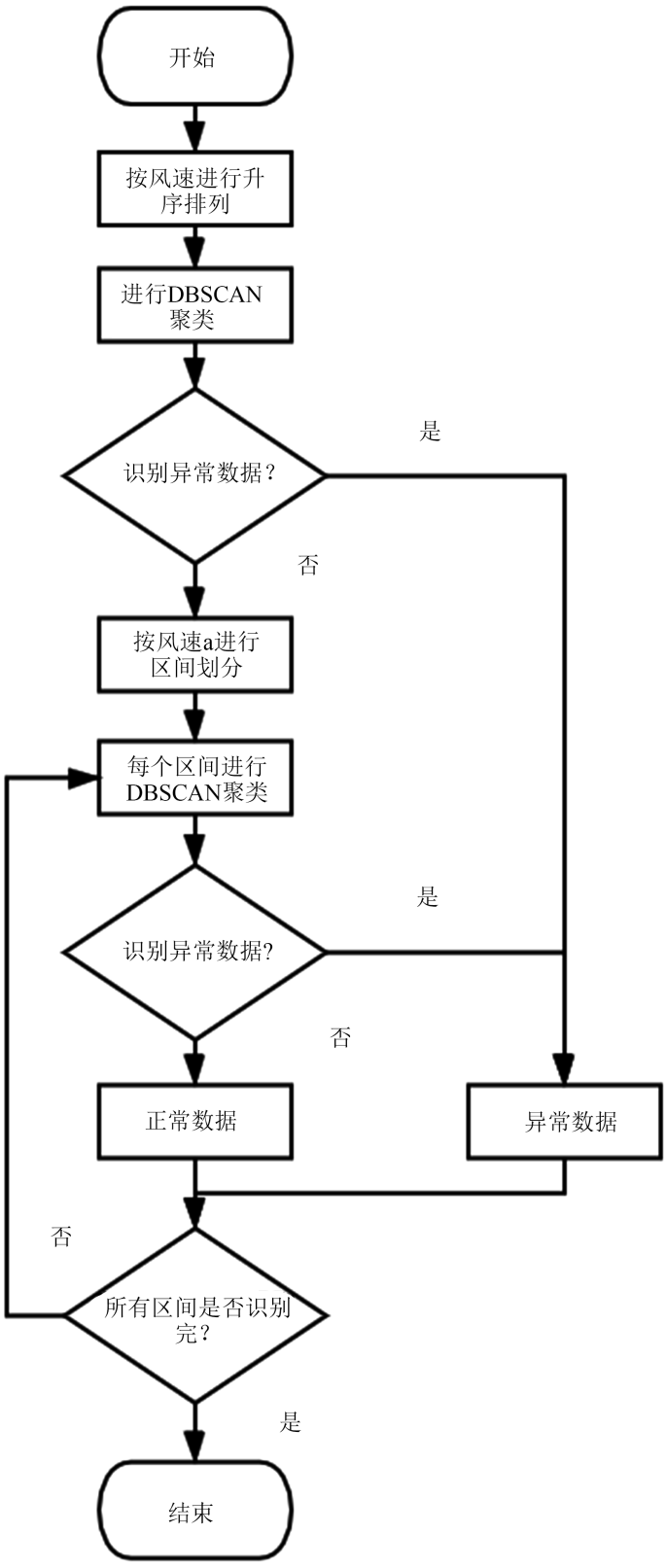
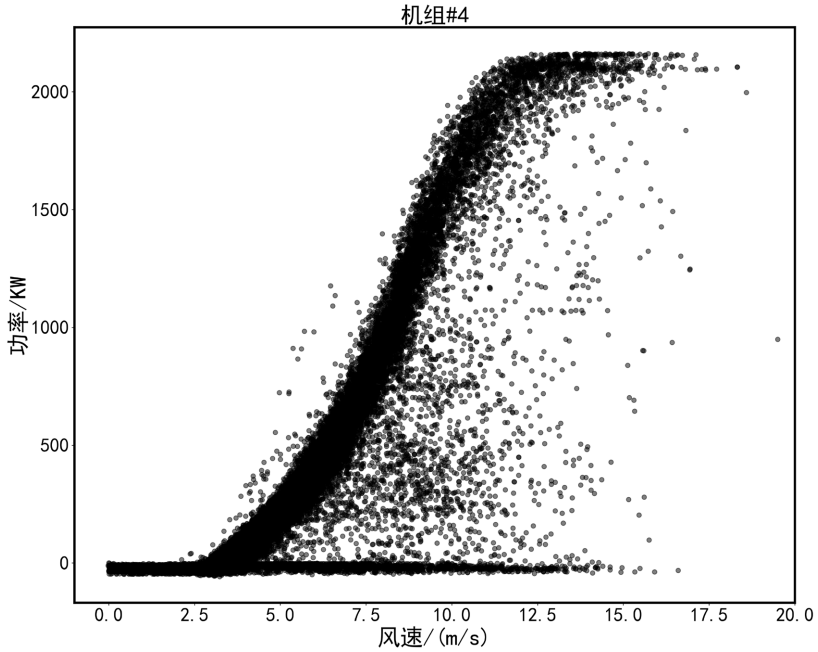
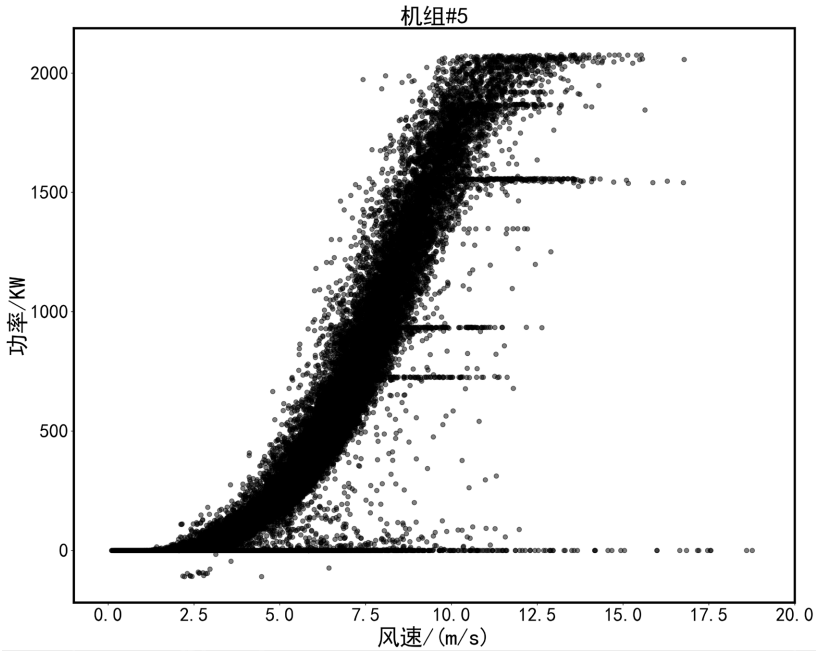



**Figure 2.** Data cleaning flow chart
**图 2.** 数据清洗流程图




(a) 4#机组




(b) 5#机组

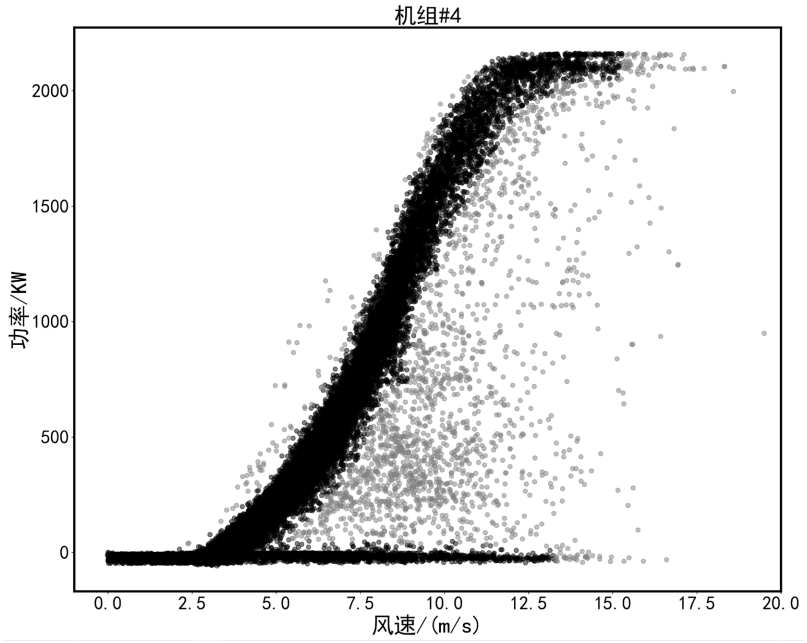
**Figure 3.** Raw data of two wind turbines: 4、5
**图 3.** 4#、5#两台机组数据样本

## 4.2. 异常数据点识别

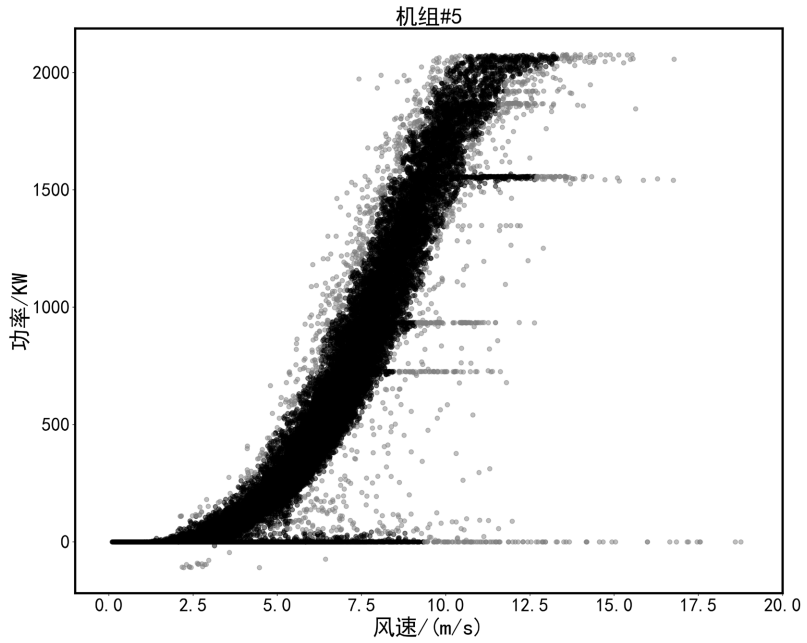
根据第 2 节提出的方法，建立异常数据识别模型，利用基于密度聚类法与基于区间的密度聚类法(Interval DBSCAN)的异常检测，对原始数据进行处理，获得清洗后的风电机组运行数据。

首先，利用基于密度的 DBSCAN 模型分别对 2 台风电机组进行聚类。通过多次合理性实验，调整最

优参数，确定出聚类半径 Eps 为 0.07，类内最小样本数 Minpts 为 16，聚类结果如图 4 所示，其中灰色圆点为聚类结果中的离群噪声点，黑色圆点表示正常数据。




(a) 4#机组




(b) 5#机组

**Figure 4.** Outlier identification of DBSCAN
**图 4.** DBSCAN 的离群点识别

由 DBSCAN 模型的离群点识别结果可知，虽然检测出大量离群异常值点，但是对于下方 0 功率堆积点、恒功率限电点，并没有很好的识别。

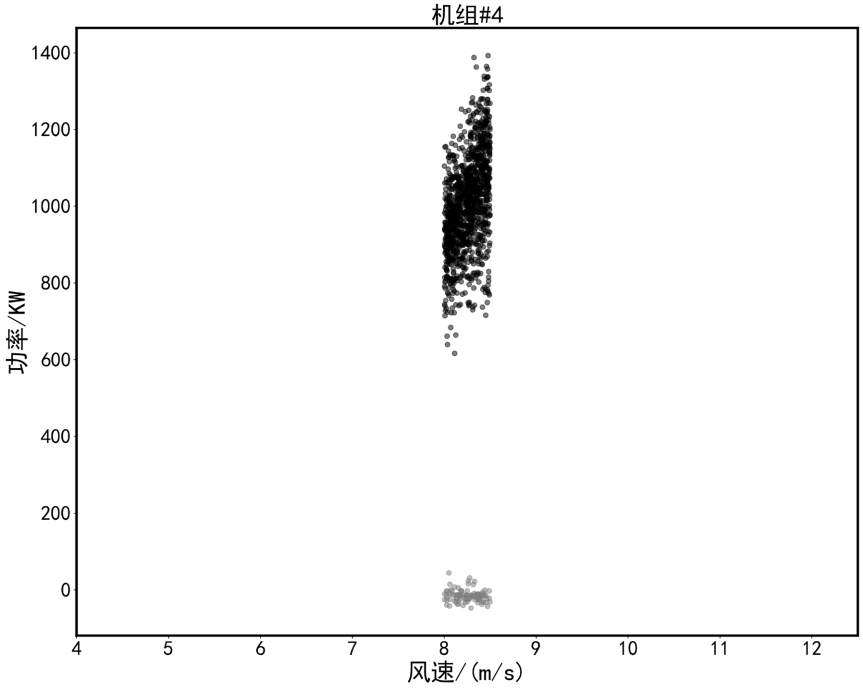
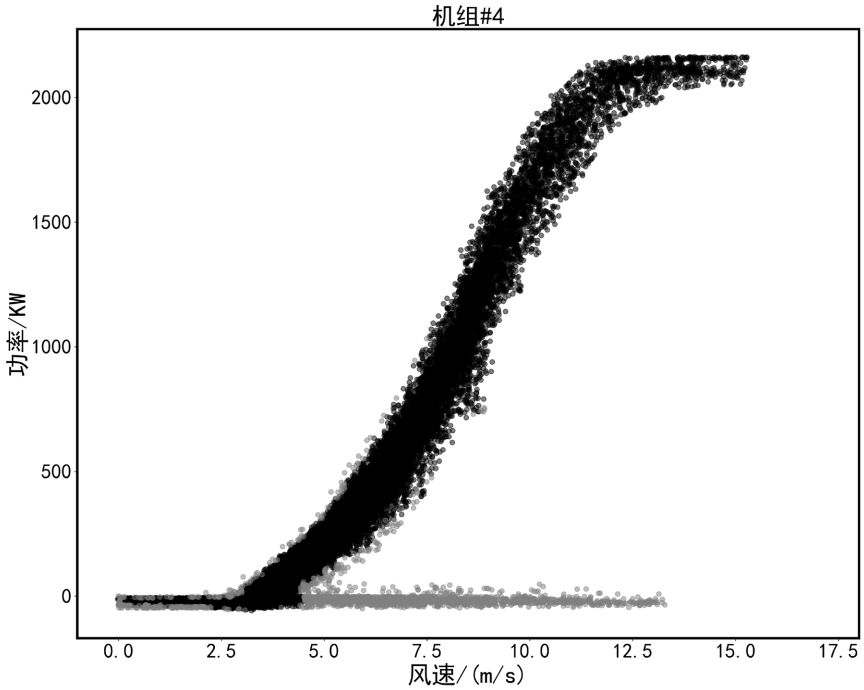
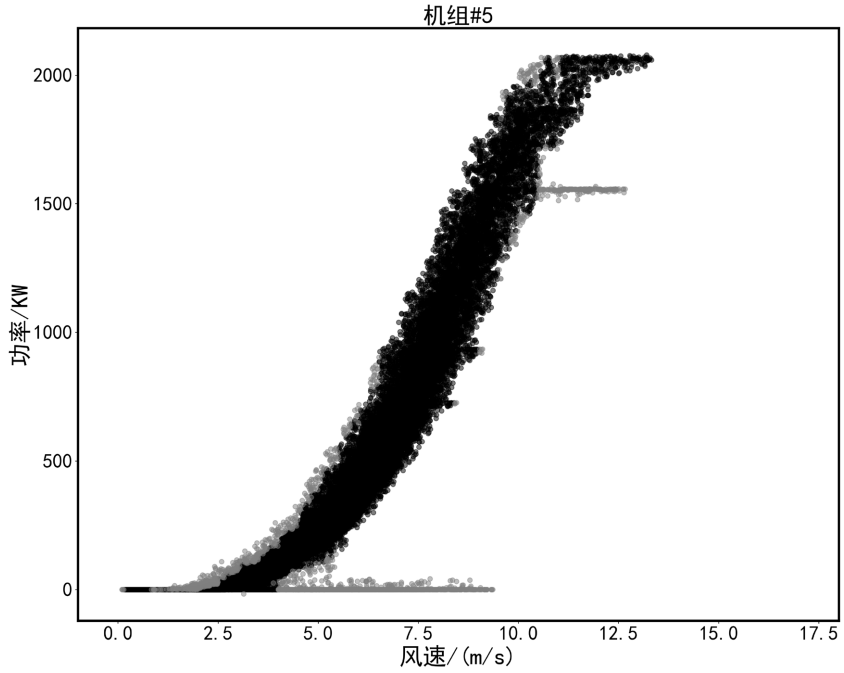



**Figure 5.** Outlier identification by region
**图 5.** 分区间的离群点识别

第二步采用改进的离群点检测模型继续进行识别，将风速按照 0.5 m/s 的间隔划分，对每个区间中P-V (功率－风速)散点图进行离群点检测。如图 5 所示，可以看到通过划分区间后，底层堆积的限电数据以及恒功率的限电数据可以很好的被识别出。图 6 为通过划分区域的离群点检测模型得到的检测结果。



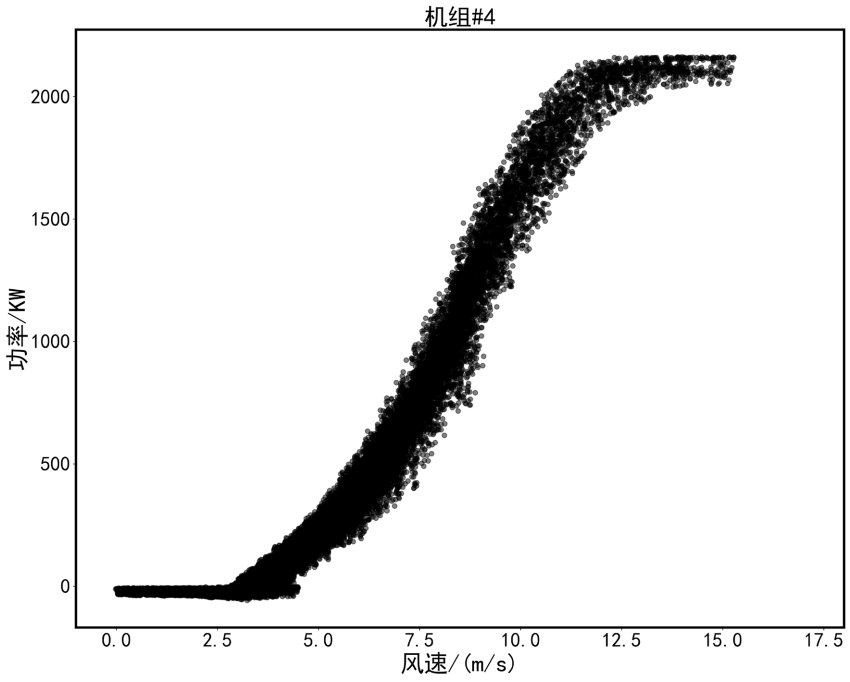

(a) 4#机组



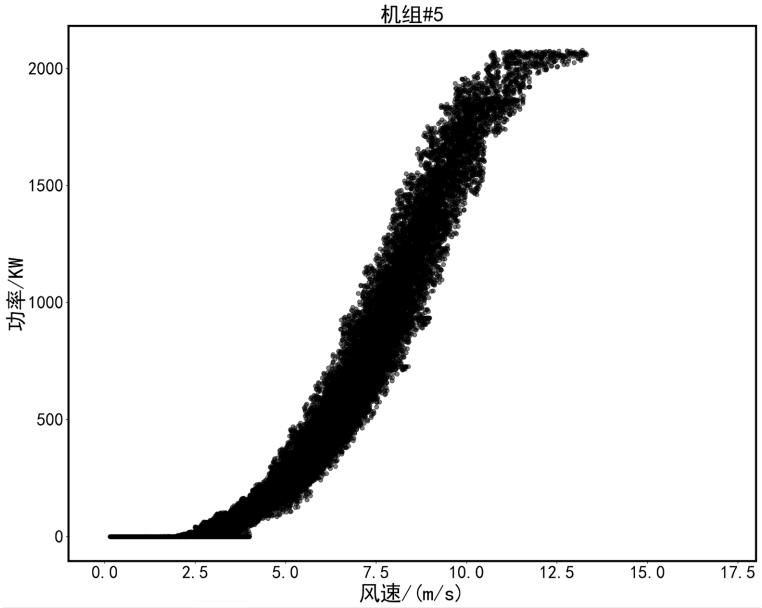

(b) 5#机组

**Figure 6.** Improved outlier identification
**图 6.** 改进的离群点识别

经过两次离群点检测与剔除后，环绕型数据以及堆积型数据都有被很好的识别，获得最终数据如图 7 所示。从图中可以看到，分散型异常数据和堆积型异常数据有明显的被剔除，可以很好的反应风速－功率曲线的特征，因此可以证明方法的实用性。




(a) 4#机组

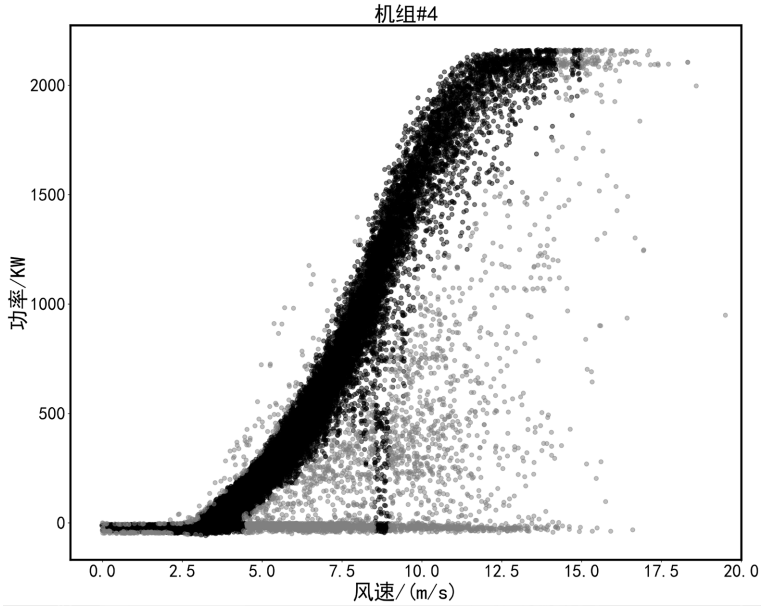



(b) 5#机组

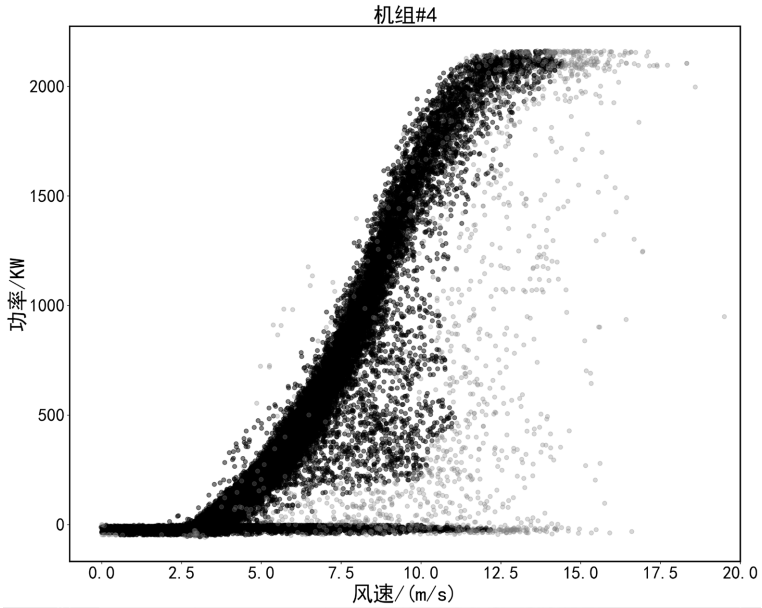
**Figure 7.** P-V scatter diagram of wind turbine after cleaning
**图 7.** 风电机组清洗后的功率－风速散点图

### 4.3. 算法对比分析

为证明 DBSCAN-Interval DBSCAN 算法在进行数据清洗过程中的合理性和有效性，本文利用 4 号风电机组，从清洗效果和异常数据检测量等维度对比分析了基于区间的 DBSCAN 以及孤立森林(Isolation Forest)异常值检测方法。两种方法对数据清洗效果图如图 8 所示。从效果上可以看到孤立森林的方法对堆积型和分散型异常数据都没有很好的识别，基于区间的 DBSCAN 算法虽然检测出较多异常数据，但是整体上还是存在漏检的问题。




(a) I_DBSCAN




(b) I_Forest

**Figure 8.** Effect comparison chart
**图 8.** 效果对比图

如表 1 所示，在数据检测率上对四种方法进行了对比，可以看到本论文的方法对异常数据的检测率是存在明显优势，在四号机组上有 8.2%的检测率，Interval -DBSCAN 有 6.7%的检测率，而 DBSCAN 和 Isolation Forest 为 4.1%和 3%，仅有其 50%左右的检测率，证明了本方法的实用性。

**Table 1.** Data cleaning effects of different methods
**表 1.** 不同算法数据清洗效果

| 清洗方案 | 机组编号 | 原始数据量 | 异常数据检测 | 清洗后数据量 | 异常数据检测率/% |
|---|---|---|---|---|---|
| D-ID | 4# | 44,335 | 3652 | 40,683 | 8.2 |
| | 5# | 50,962 | 3439 | 47,523 | 6.7 |
| DBSCAN | 4# | 44,335 | 1831 | 42,504 | 4.1 |
| | 5# | 50,962 | 1259 | 49,703 | 2.5 |
| Interval -DBSCAN | 4# | 44,335 | 2959 | 41,376 | 6.7 |
| | 5# | 50,962 | 2792 | 48,170 | 5.5 |
| Isolation Forest | 4# | 44,335 | 1331 | 43,004 | 3 |
| | 5# | 50,962 | 1529 | 49,433 | 3 |

## 5. 结论

为了解决风电机组中的异常数据对风电机组性能分析带来的负面影响，本文总结与分析了异常数据的数据类型以及产生原因，建立了两次密度聚类的异常数据识别模型进行处理。通过实验例证分析表明：

1) DBSCAN 方法可以很好地识别环绕型分散数据，第二次基于区间的 DBSCAN 方法对堆积型数据的识别效果显著。

2) 由清洗后的各机组功率－风速曲线可以看出，异常数据剔除效果好，证明了本文提出方法的合理性。

本文在基于区间的密度聚类算法的改进上进行了初步的研究，提出在各区间进行聚类算法时，考虑到划分区间后，区间数量会增多，针对调优参数过程，在依据数据的物理特性以及机组设备的运行原理的前提下，可以采用自适应确定 DBSCAN 算法参数的方法进行调优处理，这也是后面工作的研究方向。数据清洗是对风电数据进行深一步挖掘处理重要的先导工作，本文中的方法对后续的风电工作提供了良好的支持。

## 参考文献


[1] 白永秀, 鲁能, 李双媛. 双碳目标提出的背景、挑战、机遇及实现路径[J]. 中国经济评论, 2021(5): 10-13.

[2] 王一妹, 刘辉, 宋鹏, 等. 基于多阶段递进识别的风电机组异常运行数据清洗方法[J]. 可再生能源, 2020, 38(11): 1470-1476.

[3] 朱倩雯, 叶林, 赵永宁, 等. 风电场输出功率异常数据识别与重构方法研究[J]. 电力系统保护与控制, 2015, 43(3): 38-45.

[4] Wang, Y., Infield, D.G., Stephen, B., *et al*. (2014) Copula-Based Model for Wind Turbine Power Curve Outlier Rejection. *Wind Energy*, **17**, 1677-1688. https://doi.org/10.1002/we.1661

[5] 娄建楼, 胥佳, 陆恒, 等. 基于功率曲线的风电机组数据清洗算法[J]. 电力系统自动化, 2016, 40(10): 116-121.

[6] 沈小军, 付雪姣, 周冲成, 等. 风电机组风速-功率异常运行数据特征及清洗方法[J]. 电工技术学报, 2018, 33(14): 3353-61.

[7] Zheng, L., Hu, W. and Min, Y. (2015) Raw Wind Data Preprocessing: A Data-Mining Approach. *IEEE Transactions on Sustainable Energy*, **6**, 11-19. https://doi.org/10.1109/TSTE.2014.2355837

[8] 胡阳, 乔依林. 基于置信等效边界模型的风功率数据清洗方法[J]. 电力系统自动化, 2018, 42(15): 18-23+149.

[9] 范晓泉, 杜大军, 费敏锐. 风电异常测量数据智能识别方法研究[J]. 仪表技术, 2017(1): 10-14.

[10] 田书欣, 程浩忠, 曾平良, 等. 基于调频层面的风电弃风分析[J]. 电工技术学报, 2015, 30(7): 18-26.

[11] 贺玲, 吴玲达, 蔡益朝. 数据挖掘中的聚类算法综述[J]. 计算机应用研究, 2007(1): 10-13.