1. **绪论**

**1.1研究背景与意义**

根据全球风能委员会给出的《Global Wind Report 2024》，2023年全球新增风电装机容量达到117GW，比2022年增长50%，全球累计风电装机容量突破1TW。其中，陆上风电新装机容量106GW，中国新增容量达75.9GW；海上风电新增装机10.8GW，中国新增装机容量6.3GW，占比58%，全球风电行业正在迅速持续发展，中国占据了中坚地位。（1） 近年来，为了实现双碳目标，中国大力发展清洁能源，不断实现清洁能源相关的技术与管理突破，引领全球气候变化治理。

在过去的时间里，中国已经对水利能源进行了较深程度的开发，想要继续提高清洁能源在国家能源结构中的比重，就需要把开发重心移步到风能、太阳能、等自然资源，然而自然资源的利用会受到诸如成本效益、土地资源等因素的限制，于是风电建设朝着大规模区域和海上场地发展，加上我国疆域辽阔，陆上和海上风资源极其丰富，风电建设成为了重要发展方向。

风电场的智能化、规模化发展高度依赖自动化监测、数据分析与智能运维能力。大场地大容量的风电建设也带来了许多不确定的因素：大场地风况复杂，风速变化大，参与监测的设备、传感器与记录的数据数量庞大，风机故障、风速仪故障、通信故障、极端天气、奇风限电、噪声干扰等诸多因素使得风电场监控与数据采集系统（SCADA）记录的数据质量下降。而SCADA数据对风电厂风电功率预测、发电性能评估和现场设备健康状态评估等各项工作至关重要。因此，开展风电场不良数据的识别与重构研究，不仅能够提升风电场数据质量和利用效率，为智能化运维、运行优化及科学决策提供有力支撑，还能够推动风电场整体运行的安全性、可靠性和经济效益提升，具有重要的理论价值和工程应用前景。

**1.2国内外研究现状**

在对研究现状探讨之前，必须要做出声明，由于现实中风电场运行有多种因素影响，风电场长时间运行记录的数据类型多样，分布广泛，很多数据虽然脱离甚至远离特性曲线，但它们并不是由于设备故障引起，而是由人工控制或大型设备运行状态转变导致，广义来说，它们都是监测系统记录的“正常数据”，而本文中的诸如“异常数据”、“不良数据”之类的表述均包含了这些数据，这样做主要是为了方便方法设计和编程。当这些“正常数据”被识别为异常后，风电场依然可以根据工作日志和经验判断区分出这些“正常数据”，再分析其他异常来对风电场运行状况做出判断和下达设备维护指令。对这些“正常数据”重构是为了估计风电场的理想发电性能，其他指标的估计可以选择性地略过这些数据的重构。

1.2.1不良数据的识别研究现状

风功率数据的缺失和重复可根据时序筛选，易于判断。针对其他不良数据的筛选，不同的方法通常基于不同底层逻辑，其中包括有基于统计分析、基于机器人学习、基于图像识别和基于物理规则多种方法。 不同逻辑方法也可以细化分出多种种类，以（文献2） 为参考，进行分类及列举：

基于统计分析的识别方法主要有：四分位法、3sigma法、统计量法等，统计分析方法通过分析数据的分布特征，设定阈值来识别异常值，具有实现简单、计算效率高的特点，适用于数据量大且异常分布明显的场景，但对于复杂、高维度、多变模式的数据，其识别能力有限，易受极端值影响。四分位法实用性强，不依赖过往数据就能对离散群数据有较强的筛选作用，四分位法经常和其他识别方法组合使用；3sigma法要求数据服从正态分布，筛选出与平均值相差超出三倍标准差的数据标记为异常数据，3sigma法对非正态分布的数据处理效果较差；统计量法对不良数据的识别大致遵循以下步骤：首先，对数据进行预处理，剔除缺失和重复数据，然后按风速对原始数据分箱处理，对各个数据集进行功率排序，计算各个数据集的功率滑动差集（方差、标准差和它们的变化率），再人为设置滑动差阈值，将超出阈值的项及后续项划分为异常值，之前的项划分为正常值，现已演化出多种算法。

随着人工智能的发展，机器学习技术逐渐应用于不良数据识别。机器人学习的识别方法例如聚类分析（如K-means）、支持向量机（SVM）、决策树、随机森林、孤立森林（Isolation Forest）、自编码器（AutoEncoder）等，大致区分为邻近度研究技术、基于数据密度的技术和其它技术。机器学习类技术能够自动挖掘数据的内在细致结构和特征，对风电数据这种高维、非线性、不均衡数据具有较好的识别效果，但对样本量、模型参数和特征工程有一定要求。k最近邻算法(k-nearest neighbour,k-NN)通过数据之间的距离进行不良数据识别，效果良好，但需要大量数据进行训练，泛用能力不强；局部离群因子算法(local outlier factor,LOF)是基于密度的算法，常被用来识别孤立的数据点，对堆积不良数据和大型风电机组数据识别效果弱。文献

（3） 提出一种改进的密度峰值聚类算法，为了提高聚类中心和截断距离选取的准确性和可靠性，利用了四分位法和分区聚类在不同功率区间找出聚类中心，再用蝙蝠优化算法寻找最优解。该方法充分利用了正常和异常数据点之间的密度差，达到了良好识别效果。

基于图像识别的方法包括图像边缘识别技术、图像分割技术、图像像素技术。图像识别方法先对原始数据进行预处理和简单物理规则过滤，然后将功率风速散点图转化为数字图像，根据图像的结构元素，提取主要部分作为正常数据，或是由图像像素设定阈值来分割出正常数据。

另外，风电场数据往往具有强烈的时序相关性，自回归移动平均模型（ARMA/ARIMA）、卡尔曼滤波、长短时记忆网络（LSTM）、门控循环单元（GRU）等时序模型广泛应用于异常检测，时序模型通常能够模型能够捕捉数据的时序依赖关系，从而提高异常识别的准确性和鲁棒性。近年来，时序模型与SCADA多维数据相融合是一个热门的研究方向，多元数据能提高关联匹配能力，并交叉验证，更加全面地识别异常。文献（1） 中就物理机理提出一种基于特征时序波动过程匹配的异常数据辨识方法，划分时序区间的基准特征，将关键特征的时序变化与实际的风电机工况相对应，再由风电机异常工况对应出异常数据，该方法能有效识别高占比和堆积异常数据，同时适用于高分辨率数据。

1.2.2不良数据重构的研究现状

对于重复数据，大多采用直接剔除的措施。数据重构的重心在于缺失数据的补全和异常数据的替换。以下是对现有数据重构方法的简述。

利用物理特性相似的重构方法，在风电机组中寻找地理位置相近的同型号机组或联系历史数据进行参考重构，对大多数风电场来说，机组匹配难度大，历史数据不确定因素多，实施难度大，该方法的精准度和可行度较低；利用线性回归和简单插值法对数据进行重构，包括有均值插值、线性插值、拉格朗日插值、样条插值，以及多元线性回归、自回归模型、贝叶斯模型等回归和统计建模。传统插值方法效果明显，易于实现，仅对模型假设和数据分布有硬性要求，应用广泛。

利用出力相似性的机器学习重构方法和深度学习建模的重构方法，通过端到端学习，捕获数据的复杂非线性关系，实现对不良数据的高精度重构。近年来，图神经网络（GNN）等新兴模型也被应用于风电场多维数据的重构研究中。文献（5）针对功率风速特性曲线忽略环境因素，模型精度低的问题，提出一种对风能捕获状况分段的多变量风电机组功率曲线建模方法，然后基于时序匹配和四分位算法对异常数据进行重构，提高了功率曲线的模型精度，对风电场发电性能评估有较大帮助。

除了对风功率数据的处理，人们也注意到实测机舱风速数据也会收到设备等因素的影响出现缺失和异常，文献（4） 中对风速不良数据进行分析分类，提出了基于时空图降噪自编码的多点位风速重构方法，完成对风速序列的插补。文献（6） 提出一种结合信号分解的用于异常风速识别和重构的深度学习模型。

目前的理论方法在实际风电场的应用还面临诸多困难：风电场数据类型多样、来源复杂，数据间存在高维、强相关、强非线性等特点，传统方法难以兼顾效率与精度；不良数据往往具有随机性、突发性和隐蔽性，部分异常数据难以通过单一阈值或模型准确识别，方法的泛用性难以提高，高精度的方法需要计算资源的投入，难以满足工业场景下的实时处理需求。

对不良数据的处理研究方向逐渐向着深度学习、多变量数据融合建模发展，如何简化集成，提高泛用性仍是未来重要的研究目标。

1. **风电场风功率数据特性及问题分析**

**2.1风功率数据来源和特性**

风电机的风速功率数据由SCADA数据提供，SCADA数据含有多种风机运行数据，包括风速、风向、风机转速、叶片桨距角、发电机参数等，数据按时间戳记录，以便于分析长期趋势和周期性规律。功率数据主要由传感器记录，功率变送器、机舱风速仪、桨距角传感器等检测仪器每各固定时间会将实时数据传送到SCADA系统数据库，再由计算机按时间戳线索对数据产生关联，一同存储。

风电机的理论发电功率为

式中：为输出功率，为风机机舱空气密度，为扫风面积，为风能利用系数，为风速。

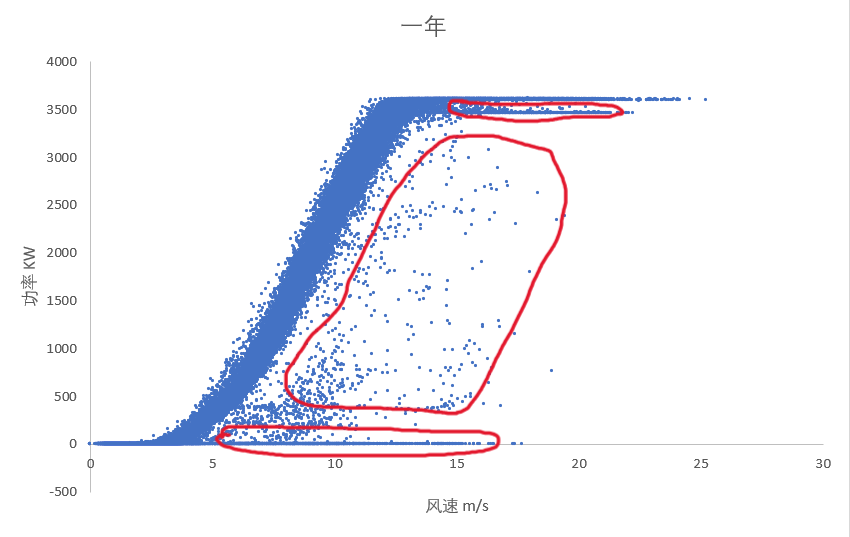
叶尖速比为叶片叶尖的线速度与风速的比值：

式中为风机转速，为风机旋转半径。

风能利用系数与叶尖速比和桨距角有关，基于贝兹理论，其理论值最大能达到约0.593，即风力发电机最大只能捕获59.3%的 风能资源。实际应用中，由于各种不可避免的能量损失，风电场风能利用系数将更降低一步。除此以外，功率数据还受到其他现实条件的约束。理论来看功率与风速的立方呈正比关系，但由于硬件技术的限制，在风机安全性控制策略下，发电功率在达到额定风速后趋于平稳；风电场所处地理位置也存在风环境复杂多变，风速波动大而迅速的特征，设备需要物理响应时间，同时设备也在运行在不同的工作状态（起动、稳定运行、停机维护等），不同工况下风机对风能的捕获效率有所不同；考虑到季节变化，不同季节电网向风电场发出的调度指令也会有所不同，会有弃风限电的现象出现。综上所述，风电机组的功率数据散点图将会出现大体数据贴近功率特性曲线，部分现实条件约束的数据点以及考虑到传感器故障、通讯故障等因素出现的错误数据。

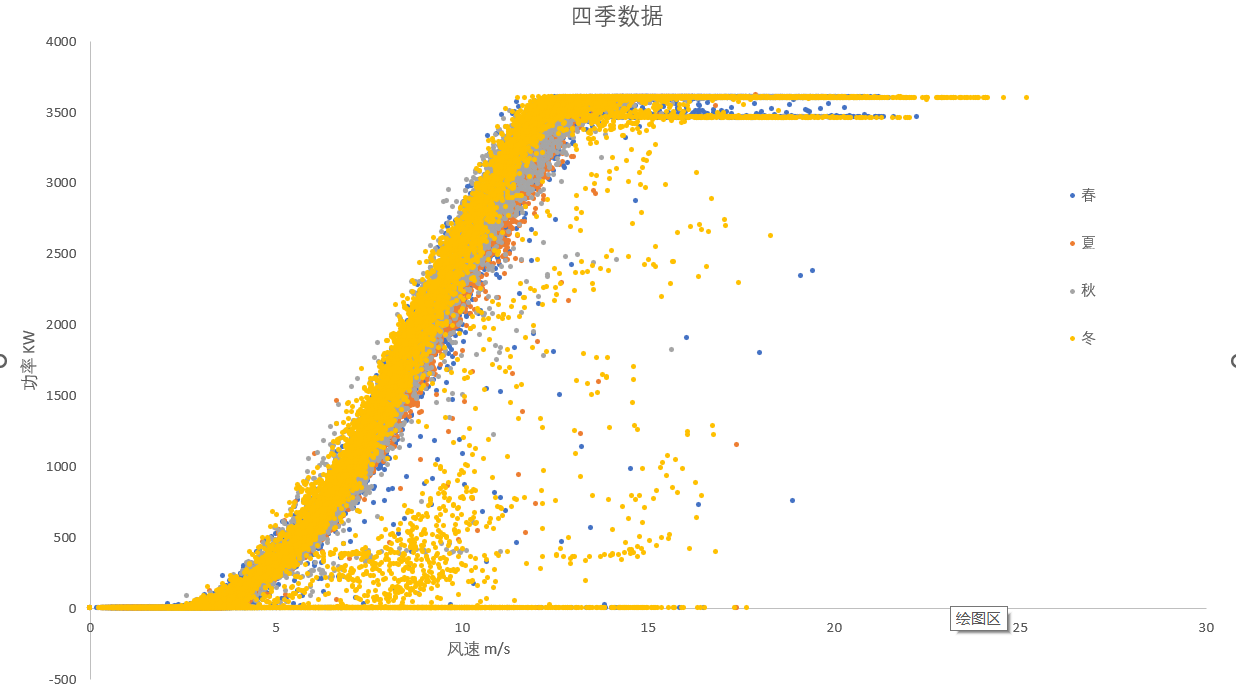
**2.2功率散点常见不良数据特性分析**

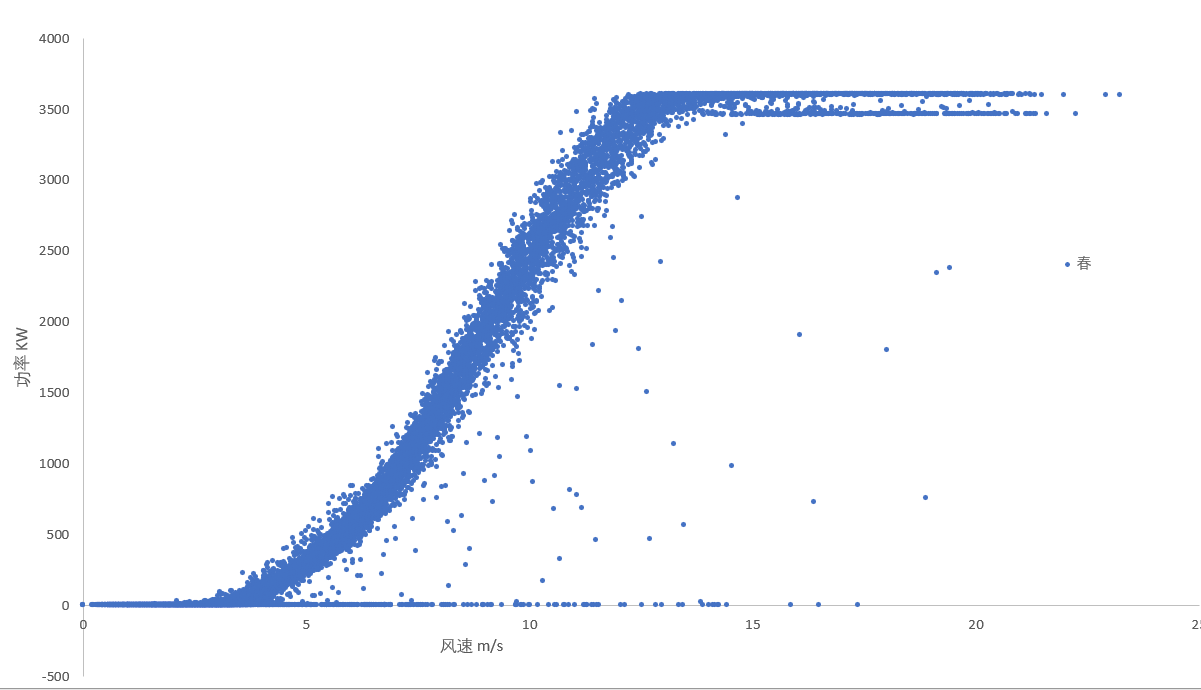
图（1） 会展示某地某风机组SCADA所记录的一年的功率风速数据，

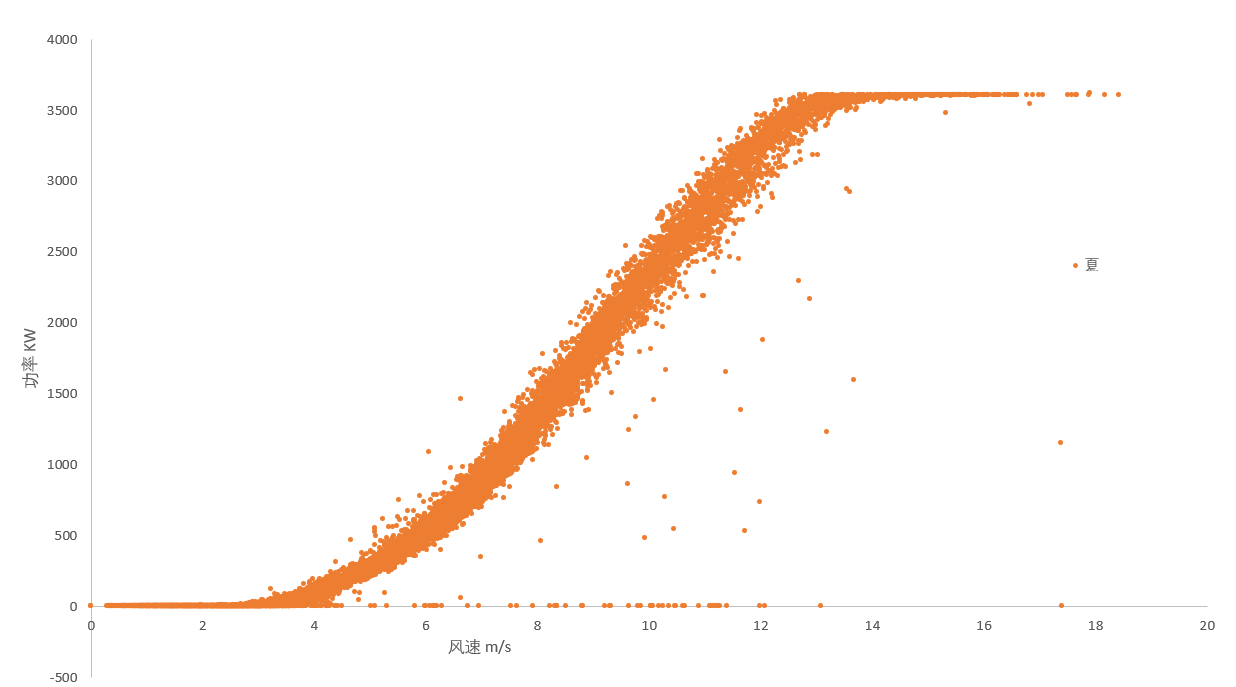


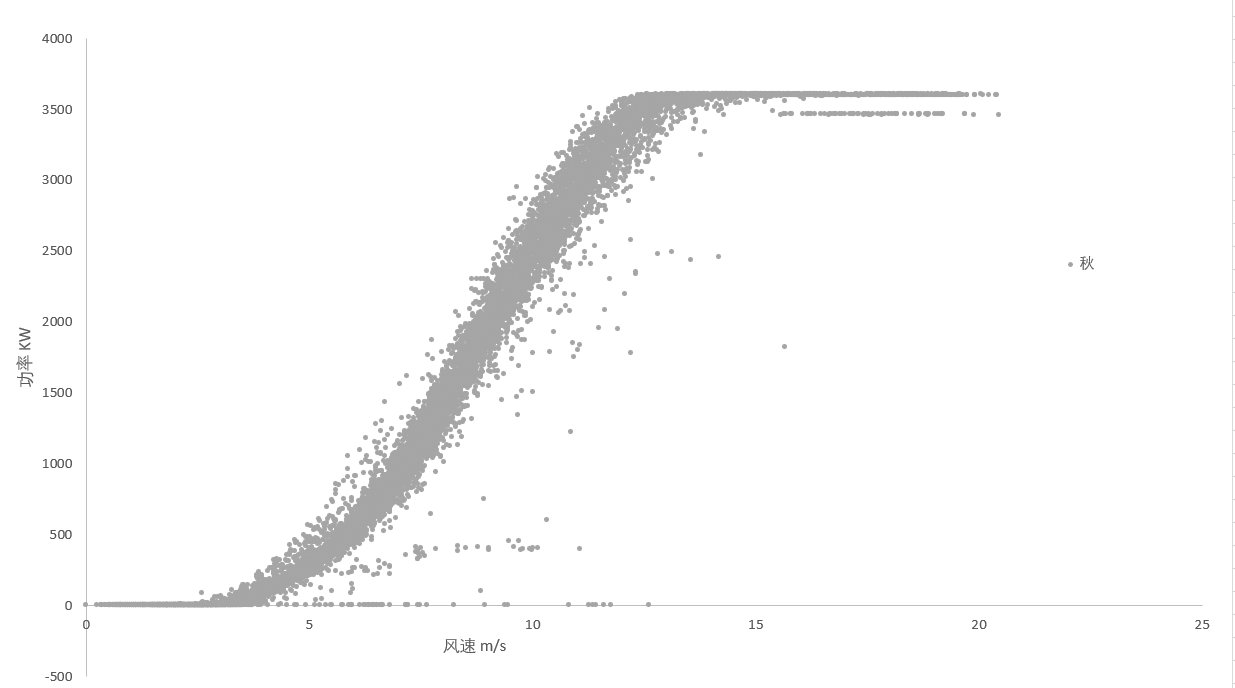
其中红线圈出的自上而下分别是不同类型的不良数据。最上方的是堆积型数据，这类数据密度较高，明显有功率一致的特点，通常是由于风电场限功率运行产生的；中间的大片数据点是离散类数据，对比堆积数据密度明显降低，在散点图上分布广泛，这通常是由于设备故障、通信等问题所致；最下方贴近0功率的堆积数据则是电机故障或停机时所记录的数据。

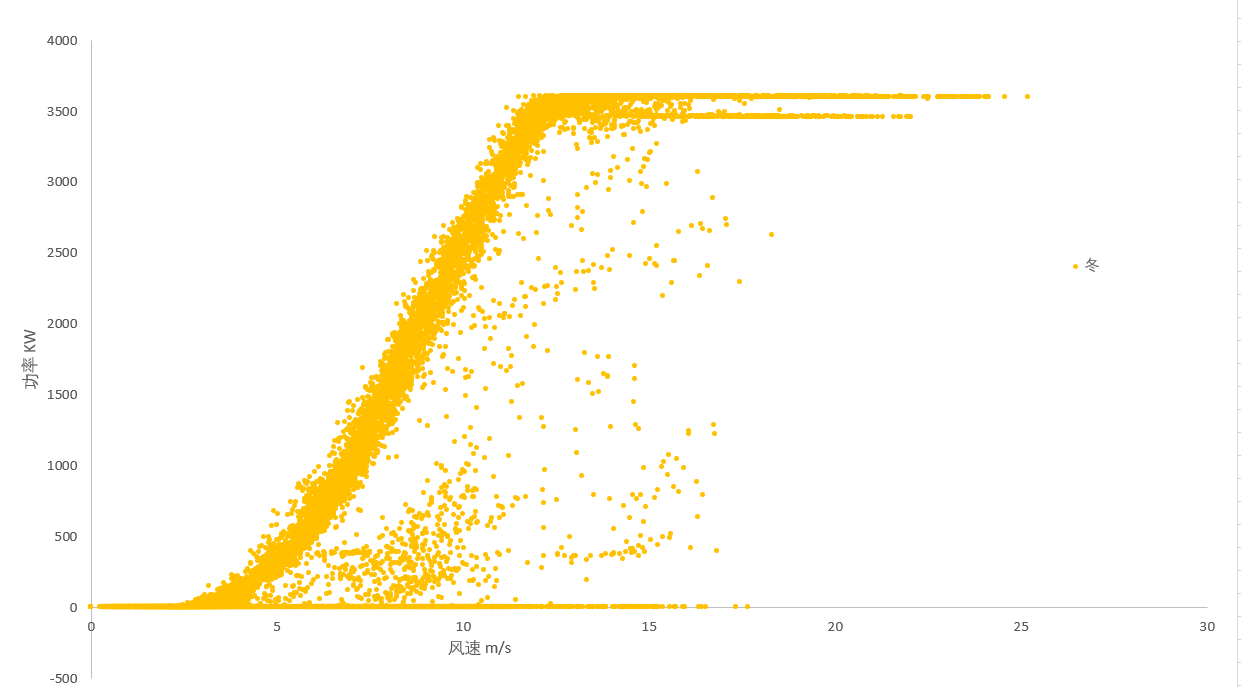
为了探讨风电数据中的异常与时间的强关联性，将一年的数据按季节划分，分别展示。











对比四季的散点数据可以看出，春季的风电数据中，有大量连续的限功率点；夏季的风电数据最为理想，整体数据基本符合功率特性，风速分布集中，没有高风速点，0功率数据么没有形成堆积，说明没有连续停机的情况，整个夏季也没有限制发电功率；秋季数据的理想程度略差于夏季，在15-20m/s的风速区间有少量限功率点，同样也没有长时间停机；冬季相较于其他季节异常分布最多，存在大量的离散异常，并且伴有大量的停机和限电情况，与理想功率特性相比有异常突出，整体数据质量较差。

针对上述的现象可以作出猜想：春季的风能资源丰富，风电场能持续高效运行，电网调度对风电场做出发电功率限制，由于春季降水频繁，低风速时会顾及设备和发电量做出停机决策；夏季风电设备和自然风况状态良好，没有极端大风天气，风速功率关系良好，曲线清晰光滑，数据质量高；秋季运行状态也比较良好，除去电网调度影响，整体数据也较为理想；冬季极端天气恶劣，风速波动大，风电机组时常出现运行不稳定的状况，异常点分布广泛且数量较多，设备调试维护和风资源利用难度大。

该风电场的例子说明了风电设备的精准监测和数据质量与环境因素有很大的关联，SCADA数据质量参差不齐，要对大量数据准确分析，需要多方面因素考虑，确保分析的全面性。

1. **基于改进风速分箱LOESS回归的不良数据识别方法**

为了让SCADA数据能更好地对风电场性能进行评估，传统的不良数据识别方法难以兼顾数据处理的准确性和处理效率，随着对这方面问题研究的深入，有了对传统方法进行改进加工的方向，结合多种方法以提高适用性。

**3.1分箱LOESS回归**

3.1.1LOESS回归原理

LOESS（Locally Estimated Scatterplot Smoothing，局部加权回归）是一种非参数回归方法，通过对数据点的某一大小的邻域进行多项式拟合来模拟复杂的非线性关系，为了能让整体曲线平滑完整，采用了数据点邻域加权回归的思想。以下对其核心步骤进行介绍。

首先，设定固定带宽或比例跨度来确定目标点及其邻域范围内的数据点，使用邻域内的数据点进行多项式拟合能较大提高与现实曲线的拟合精度，避免全局模型假设。然后为领域内的所有数据点分配权重，距离目标数据点越近，权重越大，距离越远，权重越小，在邻域选择合适的条件下，近点高权重的分配方式能提高局部模型对波动的适应性，同时降低了远点对拟合的影响。权重的计算通常使用核函数，常见的有：

三次方核：

高斯核（Gaussian Kernel）：

Epanechnikov核：

式中为目标数据点邻域点的权重，为目标数据点，为邻域点，为带宽，控制邻域的范围。

核函数中的带宽需要选取合适的值，带宽较大时，拟合的数据点多，拟合结果更加平滑，可能会忽略较多细节，弱化选用局部数据进行拟合的目的；带宽较小时，拟合结果依赖邻近点，结果更加贴合局部波动，但也容易放大噪声影响。

之后建立多项式模型对局部数据拟合，为了实用性，通常只采用低次多项式，一次多项式取为：

二次多项式取为：

式中为多项式参数。

多项式参数通常采用加权最小二乘估计，具体方法是计算加权残差平方和，求出使其达到最小值的即为多项式参数的解。

式中p为多项式阶数。

参数解的形式为：

式中是权重矩阵，是以为对角线元素的对角矩阵；是构造的设计矩阵（以二次多项式为例）：

最终预测目标点的值：

对所有的数据点进行LOESS回归，将结果组合，就能得到完整的非线性曲线。LOESS回归能灵活处理非线性数据，较好地捕捉数据的局部特性，并得出平滑的回归结果。对风电场数据进行LOESS回归能取得较好的预测效果，然而从计算原理来看，LOESS是一种效率较低的计算方式，哪怕舍弃一定对波动的敏感度，选用低阶多项式拟合，仍然需要对所有的数据点分别拟合后才能组合出结果。为了提高风电数据记录的精准和详细程度，风电场通常会会使用较高分辨率来记录数据，进而增大数据组的体量，因此，对风电场数据进行LOESS回归过程复杂且效率较低，需要改进方法。

3.1.2数据分箱处理

数据分箱是一种常用的数据处理方法，常被用来处理大规模、多维度数据，分箱的核心思想在于把数据按某种规则（等距离、同状况、时间相近、同纬度等）分为多个子集，在每个子集内进行聚合或独立分析，可有效减少噪声并提高数据的处理效率。风电场风功率数据数据量庞大，功率曲线呈现非线性，按风速对数据合理分箱，再在每个箱内执行标准LOESS流程，理想情况下，能在保证稳定拟合结果的同时减少计算量。

风电场风速数据通常保持在0～30m/s，举例按0.5m/s的间隔等距划分风速区间，将其中数据量达不到要求的风速区间标记，不再继续进行之后的步骤。完成分箱后，在每个独立的风速区间内，当目标点的邻域超出所在风速区间时，直接将超出部分的权重设置为0，这样的处理在后续步骤中降低了大量计算量。同时，风电数据高密度的特点也能保证各个区间内的拟合效果。虽然简化了计算量，但这样的处理方法也会在相邻的区间之间产生边界效应，边缘处可能会引入偏差，导致预测点在边缘处不平滑。就不良数据识别目的来看，拟合特性曲线主要是为了提供功率预测值，因此边界效应带来的后果并不突出。

**3.2动态IQR阈值设置**

四分位距（Interquartile Range，IQR）是统计学中用于衡量数据离散程度的重要指标，尤其适用于识别异常值和分析数据分布。在一组顺序分布的离散数据中，取四分之一处位置的数为第一四分位数（Q1），取二分之一位置处的数为 第二四分位数（Q2），也就是中位数，取四分之三位置处的数为第三四分位数（Q3），IQR是第三四分位数与第一四分位数的差值，即：

IQR展示了数据集中间50%的分布范围，能够排除极端值或异常值的影响。在传统的异常判断中，常按以下形式设定阈值：

3.1中对风功率数据进行了逐一预测，在每个风速区间计算设置残差的允许阈值，为了防止数据检测的灵敏度过高，设置残差允许阈值为：

借助数据分箱处理的优势，不同的风速区间有不同的数据特征，也带来了不同的阈值，更能适应短期数据波动。

**3.3简单物理规则校验**

现实中，风电场为了保证运行的效率和安全性，会为风机设定起动风速和截止风速，基于现实物理规则，需要剔除一些可能存在的异常点，其中包括：风速低于切入风速时，功率大于0的数据点；风速大于切出风速时，功率大于0的数据点；功率超出额定功率一定数值的数据点；功率为负数的数据点。以上风机参数可以根据风电机型号、风电厂技术文档、行业安全准则、设计规范等查取（常见陆上风机典型参数如表所示），或者结合SCADA数据大致判断。

**3.4数据测试与结果分析**

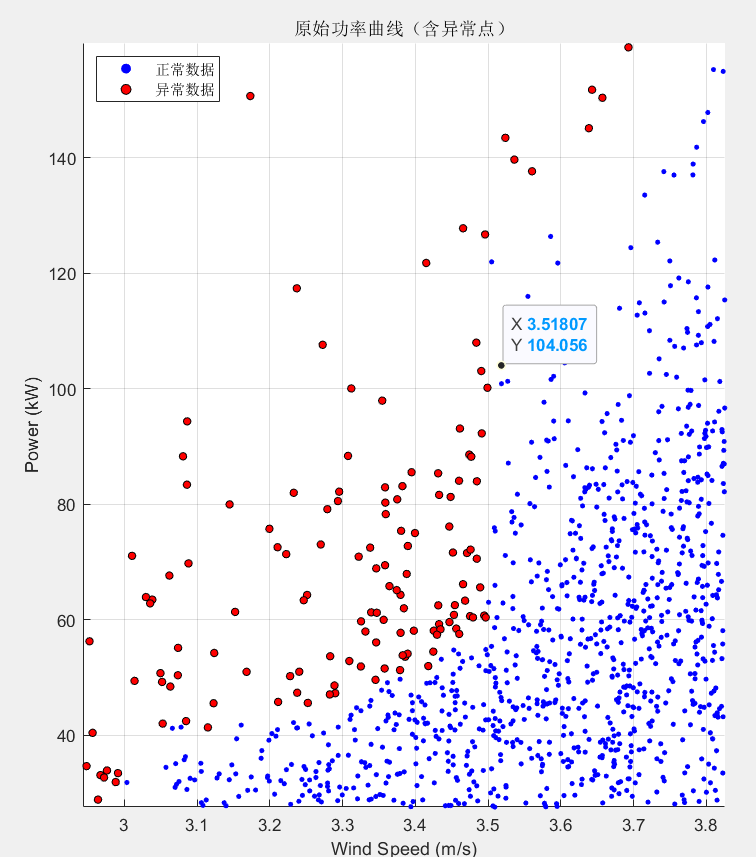
本次测试的数据来自Kaggle网站的公开数据集“Wind Turbine Power Data”，其中包含某某风电机组2018年整年的SCADA数据，采用10分钟分辨率共计50530条数据，记录风电机组包括时间，发电功率、风速、该风速下的理论功率、风向几个维度的数据。在测试中使用其中时间戳、功率、风速进行测试。

3.4.1样本数据测试分析

由于原始的风电数据中存在大量脱离特性曲线的数据点，现对原始数据使用程序，查看测试结果，并对结果进行分析。将数据带入matlab程序，默认参数风速分箱宽度为0.5m/s,平滑系数取0.25，动态阈值取为1.5倍iqr加上残差绝对值的中位数。运行得到效果图： 现就测试结果进行分析。

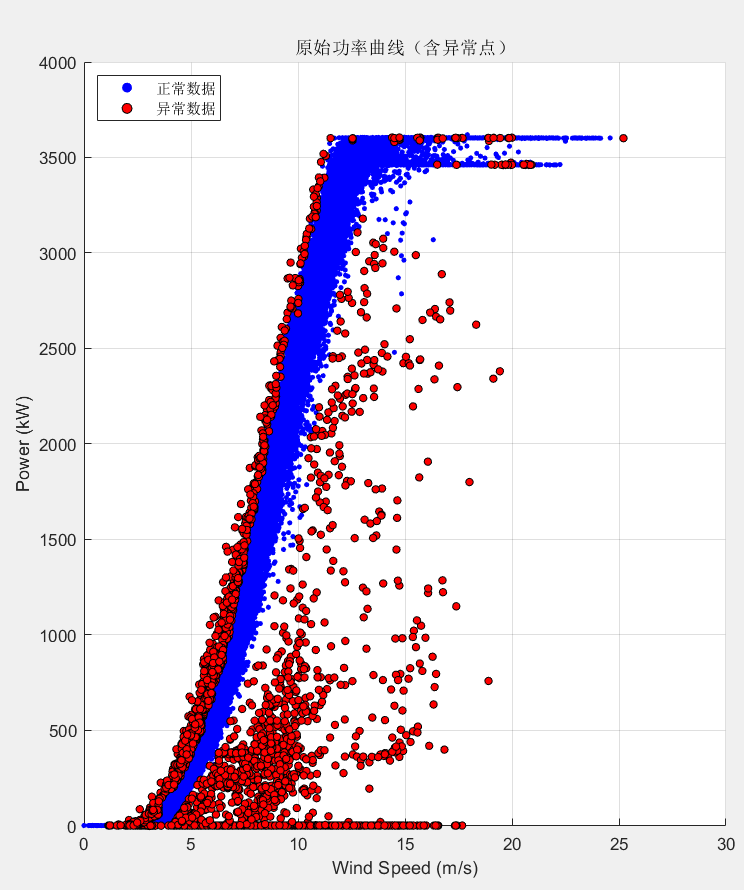
结果显示程序共标记5102个红色不良数据，从图中可以看出，程序对不良数据的识别有不错的效果，针对大部分离散型异常和底部堆积型异常数据都做出了有效识别，效果良好。此外，也展示了识别方法的不足，逐一进行分析。

在功率曲线呈上升的部分，程序将很多功率数值较高，超过额定功率区间的点识别为异常数据，在风速低于10m/s的部分分布最多，且紧贴正常数据，根据LOESS回归预测的原理可知，某一点的预测值受到邻近数据的影响且权重较大，当某点风速箱内低功率异常点很多时，会使得预测值偏低，导致将一些比较正常的数据划为异常值产生误判，在动态阈值设置中曾对此类异常误判有过预防，如式13中添加了残差绝对值的中位数作为阈值设置标准，如果某风速箱中残差分布散乱且数值较大，那么自动设置的允许阈值也会比较大，进而对正常数据的过敏起到一定缓冲作用，不过任然无法完全杜绝异常误判，图中10m/s以下数据中存在大量离散的低功率异常值和底部堆积，可能造成了部分异常误判。放大识别结果来看，如前文分箱处理解析中提到的，在风速箱的边缘位置出现了明显的边界效应，如图 所示。在3.5m/s风速附近，正常数据与异常数据的判定有明显的分界线。



在特性曲线接近额定风速的部分，程序对一些不良数据点没有做出标记，识别效果不够好。在风速接近额定风速时，风电机组的输出功率理论上趋于额定功率，功率不再随着风速快速上升，功率曲线变得平坦，风速之间的功率差异明显减小，正常点和部分异常点（如限功、滞后、短时波动等）在数值上相差不大。这时LOESS拟合的残差变小，异常点与正常点的差距变小。同时，在残差阈值设置时还添加了“残差绝对值的中位数”这一项，使得阈值检测更加宽松，难以分辨轻微异常，异常检测的灵敏度下降，这体现了数据分布特性与程序检测机制之间存在一定冲突。另外，由于风速分布一般呈现偏态，额定风速附近的风速箱内数据会少一些，这可能也导致了局部拟合的稳定性下降。

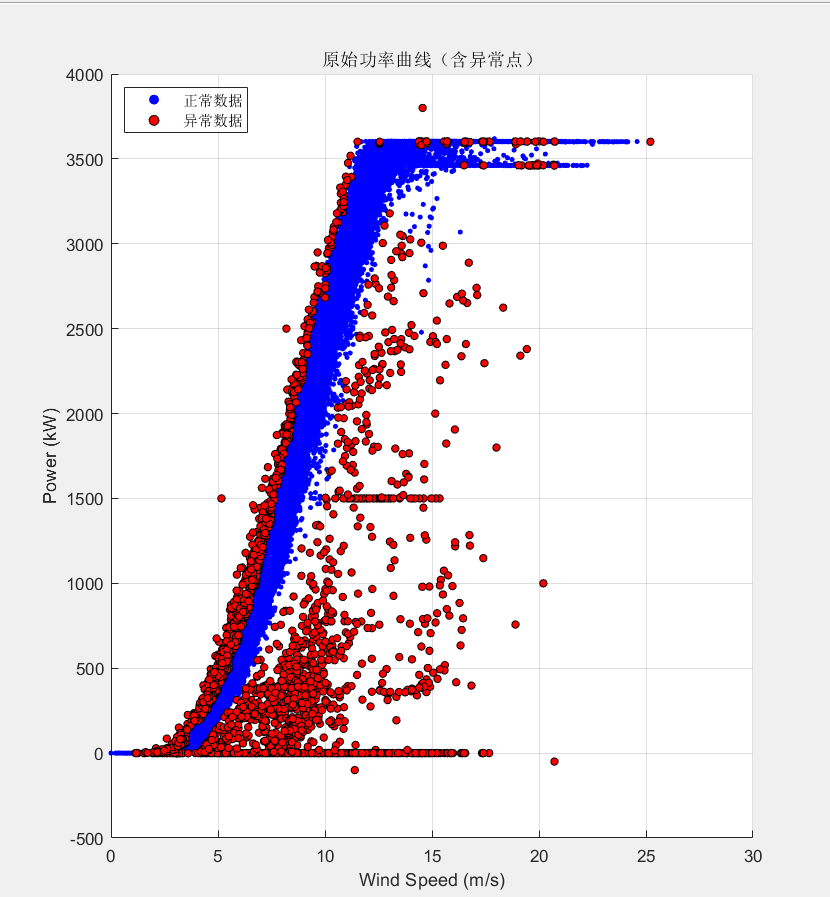
此外，程序在对接近额定功率的堆积异常点识别能力不足，甚至存在额定功率点被判为异常。在风速大于额定风速时，理论上功率应该固定在额定功率附近，而网站提供的风电数据来看，该风电机组一年的时间内存在大量限电运行的工作状况，且如图中所示，该风电厂限制运行功率接近额定功率，额定功率点与限功点数值相差不大。风速大于额定值以后，LOESS拟合曲线本身接近一条直线，残差分布较为集中，IQR数值小，阈值的设定很大程度由残差绝对值的中位数决定，容易导致阈值设定过窄/过宽的极端情况：偶尔额定点与限功点全部符合允许范围，偶尔两者都被判为异常点。很多异常点并没有从数值上显著偏离正常点，同样难以被识别。综上，程序在高于额定风速的区间识别能力有限，在异常点与正常点的统计特征接近，数值相近时表现较差，可能需要引入更细致的工况分类和针对性的阈值设定才好应对这种情况。



3.4.2人工异常引入测试

为了更加全面地验证设计方法对不良数据的识别能力，在对原始风电数据进行了识别测试以后，额外引入人工异常点来进一步测试性能。针对风电数据可能会出现的异常，更改部分正常数据的功率值，其中有：负功率点（设置散点图坐标（11.39，-100）（20.72，-50））、数据量较少的高风速低功率点（设置散点图坐标（15.15，2000）（20.2，1000））、功率过高的点（设置散点图坐标（8.192，2500）（5.16，1500）（14.56，3800））、远离额定功率的限功点（将它们的功率限制为1500千瓦，风速范围约为9.4-15.3m/s）。

将更改后的数据重新带入程序运行，结果展示如图 所示。



从结果可以看出，设置的多个异常数据均被程序做出标识，证明了程序对远离额定功率的堆积型异常数据有一定的识别能力，总体识别效果良好。

3.4.3方法参数调整

在运行不良数据识别程序的过程中算法参数对于异常检测的灵敏度、准确性有直接影响，通过调整参数可以改变程序对不同类型异常数据的识别效果。在本文提出的不良数据识别方法中，主要可调节的参数有：分箱宽度、LOESS平滑系数span、残差的允许阈值设置、分箱基本样本数、风电机相关参数。

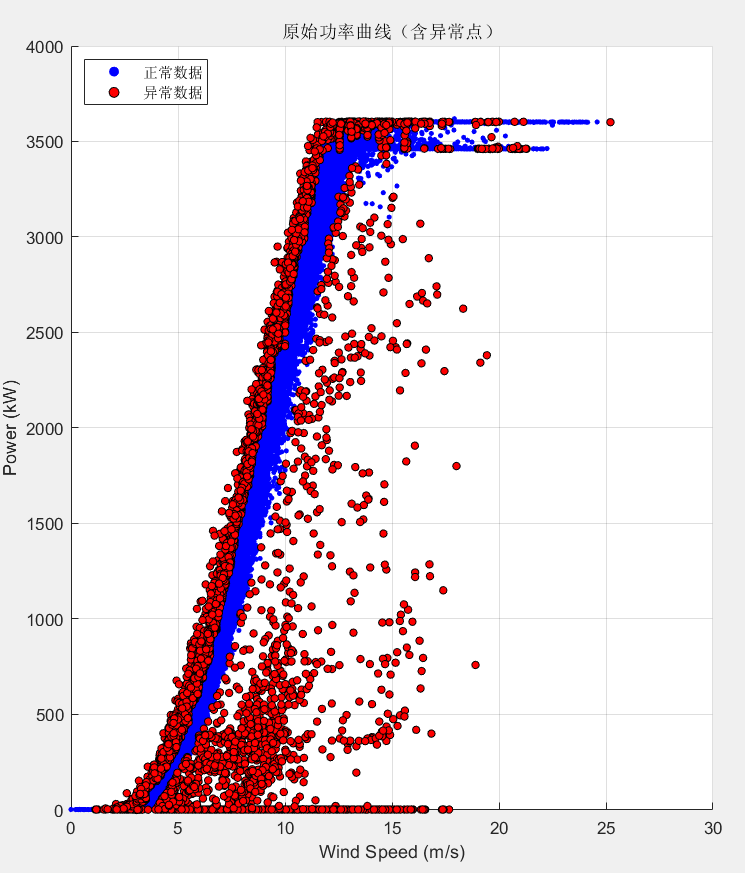
其中分箱宽度主要影响预测值的精确性，分箱越多，局部拟合越精细，能更好适应风速功率的非线性关系，也能更加细致地划分判断阈值，同时，分箱数量多意味着用核函数计算地权重少，计算量少，程序的响应速度快，风速箱的边界效应可能会有凸显，也意味着风速箱内样本数量减少，数据密度不高的位置可能会出现局部统计不稳定，容易误判或漏判；而如果分箱数量少，可能导致局部特性被忽略，异常点被掩盖，灵敏度降低。理想状况下，应该根据数据特性人工调整不同风速的分箱宽度，数据量大时可适度减小分箱宽度提高识别灵敏度，数据量少或稀疏时适当加宽避免误判和漏判。

平滑系数的设置和数据的噪声程度有关，平滑系数越小，预测曲线更贴近原始数据，更容易突出异常点与预测点之间的差距，使其更容易被识别，但相应的，噪声干扰对曲线拟合的影响也会被放大，过拟合噪声很容易丧失程序的识别能力，为了达成异常数据识别的最初目标，在不确定数据噪声程度的情况下，平滑系数不宜设置得太小。

残差的允许阈值设置是在异常数据识别的严格程度和减少正常数据被误判之间进行取舍，通常要根据数据识别和重构的要求和目的来进行调整。现将允许阈值设置为1.5倍iqr，就原始数据运行程序，结果如图 所示，与3.4.1中默认参数设置对比，共计标记不良数据8894个，缩小残差的允许阈值明显提高了程序的敏感度，但从正常数据的角度来看，也放大了过敏程度。除了根据需求和目的性强弱来调整，也可以适当参考重构结果来调整，详细分析在下一章节中会进行说明。

分箱基本样本数即是允许通过分箱流程进入下一回归步骤的最小样本数，是为了避免风速箱中样本数量太少，使得后续回归识别参考意义低，浪费算力资源的情况。分箱基本样本数不应设置过高，这是因为许多风电场风速分布呈偏态，极大风速区数据较少，如果分箱样本要求数太高，很可能会跳过极大风速区，相关的异常难以被识别，提高总体样本量也有利于避免这种情况发生。

风电机相关参数如切入切出风速、额定功率等最好是参阅设备上的标准数值直接设置，假如没有，也可以根据数据分布的明显特征进行估计，在有大量数据支撑的情况下，估计值与实际值不会有太大偏差。



1. **不良数据重构方法设计**

上一章的内容中对风功率特性中的异常点进行了筛选。在风机运行的过程中，对某些特性明显的数据集进行分析，有经验的工作者们能对风机的状况得出判断，进而跟进维护修理或运行调整。除去识别，对这些异常点进行重构矫正也是一项很重要的工作。按理想运行状况重构出的风功率数据能够用于风电场地理位置风能可利用资源的评定；按照风机实际运行工况来进行重构矫正的风功率数据能用于该风电场某一时段的综合发电性能评估。

**4.1物理规则修正**

根据物理规则和风机安全运行准则，风电数据应该具有以下特征：在风速低于切入风速时，发电效率低，桨距角设置为90°，系统不运行，发电功率等于0；在风速高于切出风速时，发电机组被切出电网，发电功率等于0；在默认风机没有改装的情况下，受到风能利用系数的限制，不存在功率远高于其它相邻风速点的数据；不存在数值为负数的功率数据；功率数据不超过风机的额定功率。

与上述特征不相符的数据可以直接划定为错误数据，它们可能是由各种因素造成的，比如传感器信息传输出现断触延迟而导致系统记录的功率与风速相互之间不匹配、仪器线路接触不良等。

对于风速低于切入风速、高于切出风速的异常点，需要置零处理；出现功率远高于其它相邻风速点的数据时，要区分孤立点和聚集点，对其中孤立的点回归重构，而大量的聚集点通常是由于风速仪出现异常导致的，需要进行功率数据迁移，使其附和功率特性曲线；其它异常点继续尝试回归重构，对重构后的结果，同样要施加物理规则约束，以保证其符合实际情况要求。

**4.2风速邻近点线性回归插值**

4.2.1滑动窗口检测

在对一组存在特定特征的数据中出现的一个异常点进行重构时，我们通常会选取异常点相邻的少数数据进行预测，假如曲线呈现明显地单调变化，那么仅需要异常点两侧的数据就能划定重构范围，获得良好的清洗效果。但在重构风功率不良数据点时，整组中存在大量异常点，且堆积型的异常数据会使得部分不良数据相隔距离小，甚至相邻。如果仅取异常点周围的少量数据用作重构，异常数据之间很可能会相互影响，无法正确重构。

为了方便程序化，选择为异常点划定数据窗口，在筛选参与重构的数据环节，需要计算邻近点与重构点的风速差值，并排序取最小的一部分点，滑动数据窗口能省去大量数据计算，提高了运行效率。以异常点为中心，向两侧搜索一定数量数据点，然后剔除其中的异常点，检测剩余数据的数量，若满足了一定的数量条件，即可利用这些邻近数据进行重构；若数据窗口剔除的数据过多，剩下的数据重构结果可能不够精确，因此需要扩大数据窗口，使得参与重构的数据点数量足够且距离重构点足够近。

4.2.2线性回归插值

有了有效的邻近数据，考虑基于最小二乘法构建线性回归模型。设定功率与风速的关系为：

式中代表功率作为响应变量，为设计矩阵，设计矩阵的第一列全为1，第二列为预测变量风速，为待估计的系数向量，为服从正态分布的误差项（）。

构建模型的核心目标在于求解最小化残差平方和：

可以化出正规方程：

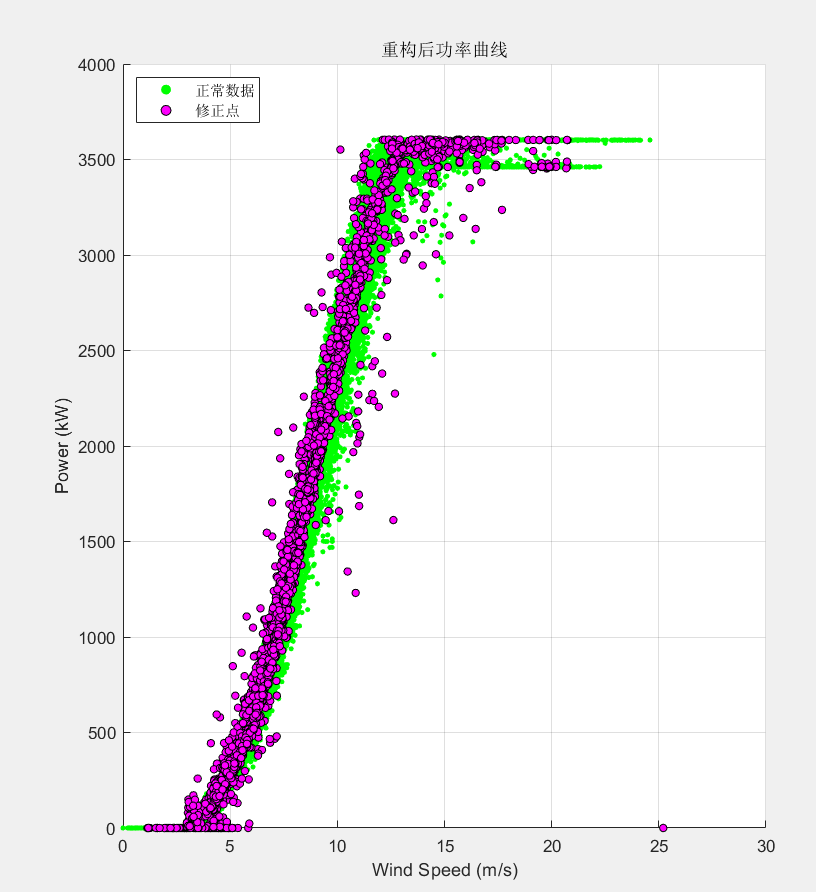
假设可逆，即可解得：

根据模型对所有的异常点进行回归插值以后，再次对重构点进行物理规则检验，以保证数据不会突破物理规则限制。对不能通过检验的点进行置零安全回退，在输出结果时，异常零点可以给予使用者提示，进而根据要求判断数据重构的可用性。

**4.3数据测试与分析**

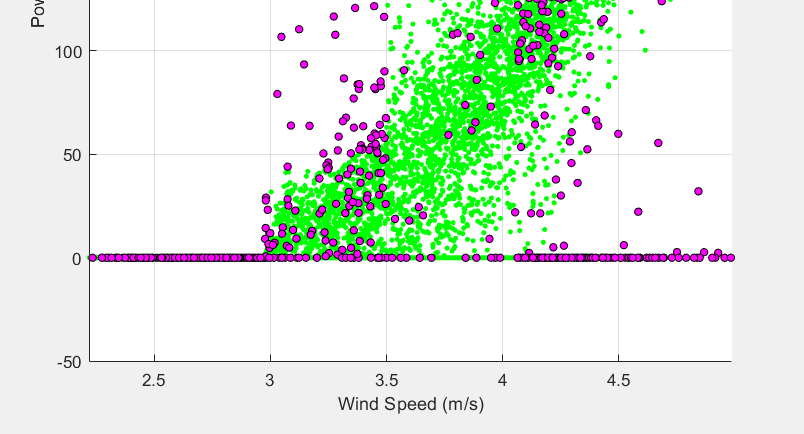
4.3.1人工引入异常数据测试

为了更全面地展示重构方法的数据重构效果，直接对3.4.2中引入了人工异常的数据进行重构，默认设定数据窗口大小为12，即数据点前后各取6个数据的范围，参与回归拟合的邻近点不超过10个。带入程序运行，数据重构的效果由图 所展示。



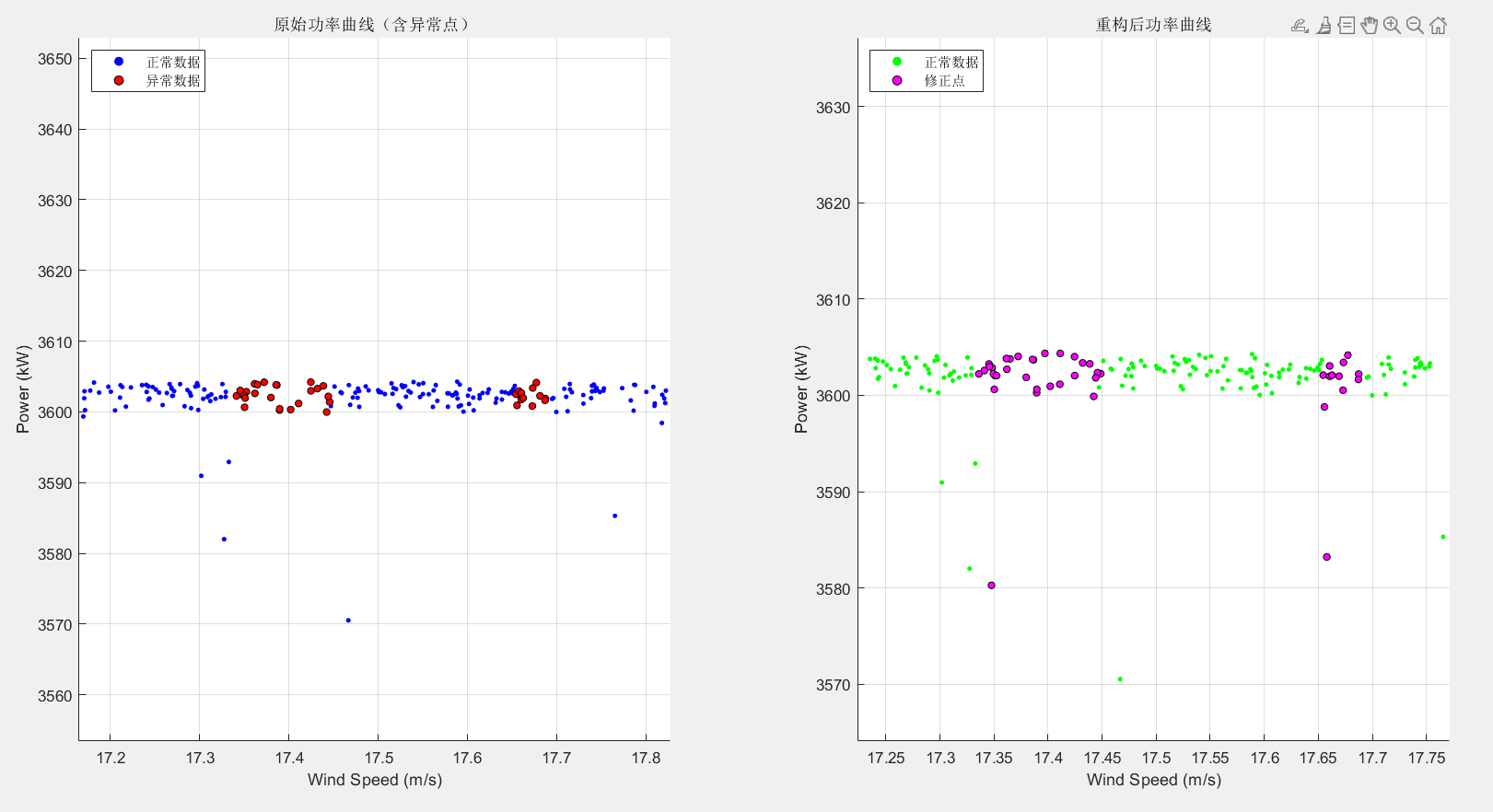
图中粉色的点代表重构数据，重构后的功率数据呈现出更加规整，符合物理规律的S型曲线，人工引入的异常数据得到了合适修正。重构点与正常数据趋势接近，整体数据更加紧凑，符合风电机典型功率特性曲线，整体重构效果良好。

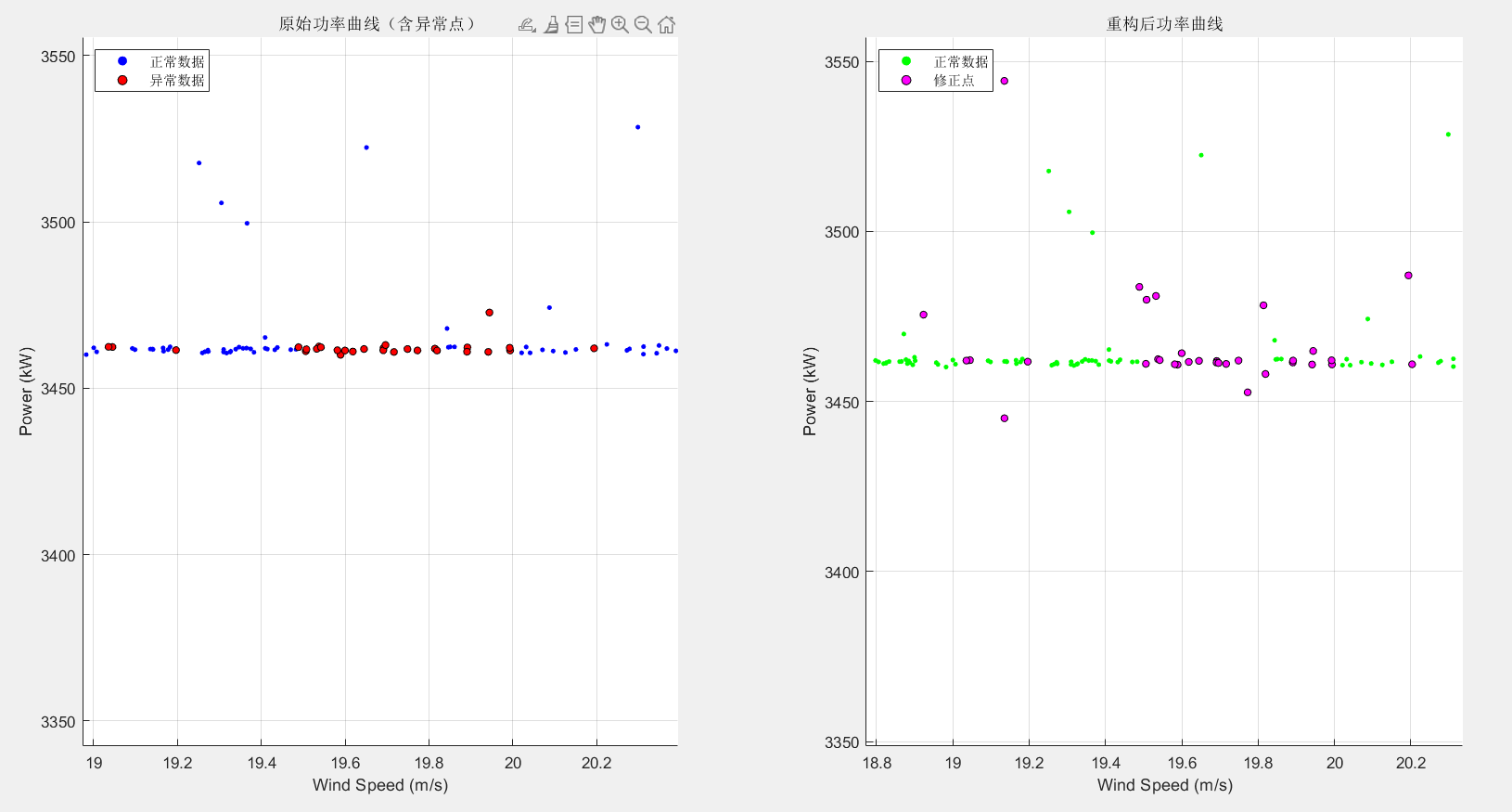
在低风速区域，正常数据与重构数据紧密重合，程序将低于切入风速的点强制置0，导致切入风速处出现锯齿 ，锯齿的大小与切入风速的设定和估计有关，根据数据点的分布减小切入风速的设定可以减小锯齿的大小，使得图像更加平滑。而由于切入风速附近的数据点功率接近0，对整体数据影响也不大。在风速约4-5m/s的区域内重构的数据出现了零点，结合重构方法的原理可以得知这些数据是由于回归预测达到了负功率，触发了程序的安全回退置0。本文采用的回归插值方法是普通最小二乘法，低风速时功率数据接近0，难以避免重构出部分负功率数据。



在功率上升的风速范围内，重构数据出现了少量逸散数据，也是由重构方法所致：本文方法中选取回归模型为简单直线方程，主要目的是为了更好地捕捉数据的变化趋势，实现更加准确的拟合效果，而这种方法容易数据的大幅度波动的影响，使得少量数据重构效果差；另外异常数据没有被识别出来也导致了重构性能减弱，例如在额定风速附近，第三章中提到此处不良数据的识别效果较差，有少量低功率数据没有被识别出来，使得正常数据的功率分布范围很大，当这些低功率数据附近的点进行重构时，滑动窗口搜索到了这些低功率数据，并使用这些数据代入回归模型，导致重构结果不合理。好在逸散数据数量很少，也并没有完全脱离特性曲线主体，对数据的整体评估不会起到太大影响。

在高风速范围内，对比异常数据分布与重构数据分布，图 和图 分别是第三章数据分析中提到的被误判的额定功率数据和被识别出的邻额定功率限功数据的识别重构对比图，从图中可以很明显地看出，这两种数据重构效果非常不明显，重构前后数据点位置几乎没有变动，这也是对堆积平台采用数据窗口回归带来的特性。上述分析表明，不良数据识别方法中对正常数据的误判在合适的重构方法作用下不会产生严重的后果，在只追求重构结果的情况，可以适当放宽对正常数据误判的要求。因此可以对3.4.3中的内容进行补充：在只追求重构结果的条件下，可以适当提高设置阈值的要求以追求更高的识别灵敏度，对正常数据的误判不会对它们造成过于严重的影响。





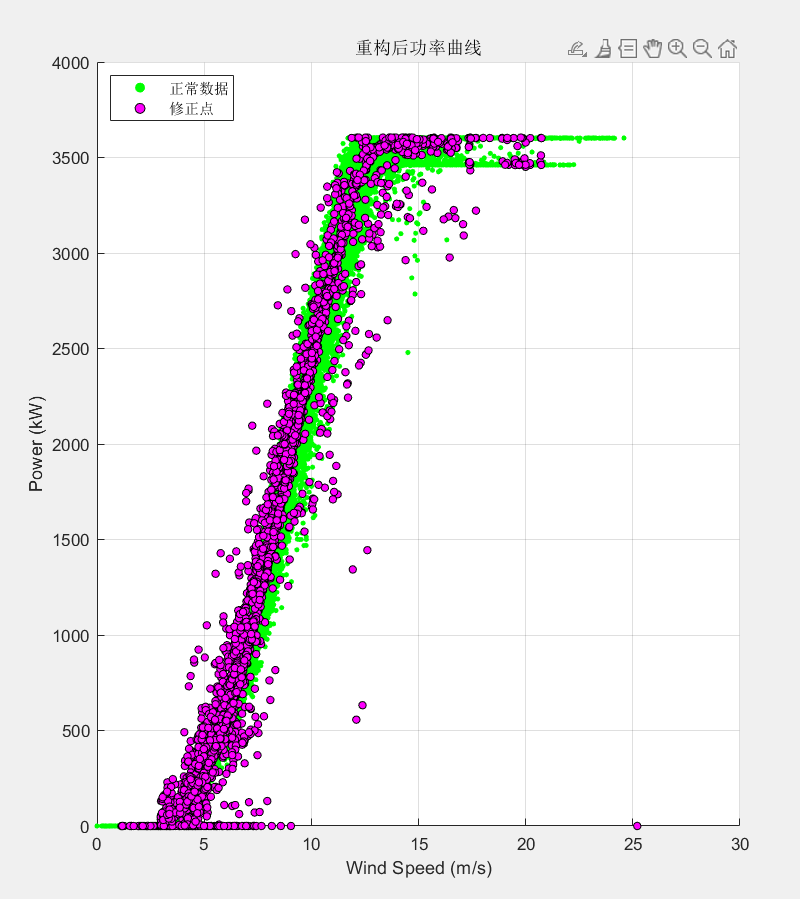
4.3.2数据重构方法参数调整

有关重构环节涉及的参数调节主要包括滑动窗口大小、判断是否需要扩大搜索区域的正常点数阈值和最多选用参与重构的邻近点数量。

滑动窗口的设置主要是为了减少计算量，减轻运行过程冗杂程度，同时参与曲线拟合的数据均为邻近重构点的数据，只要能够容纳足够的邻近数据，原则上窗口大小可以紧小设置。考虑到异常数据的分布特性，在异常数据堆积紧密的特殊情况下，会出现重构点附近异常点过多的情况，假如窗口大小设置地很小，数据窗口中可以参与回归模型的正常数据太少，可能无法支撑重构结果的准确性；而如果是对突变型的孤立异常点重构，它的附近都是可参与回归模型的正常数据，那么小窗口也能保证重构的准确性。为了同时维持小窗口的灵活性和应对上述特殊情况的能力，另外设定参数，检测小窗口内正常数据的个数，如果正常数据低于一定阈值，则直接放弃窗口搜索限制，检测距离稍远的正常数据，风速差距大也同时意味着局部特性的捕捉能力下降，重构结果更加倾向于保证贴近功率特性曲线。设定的扩大搜索的正常点阈值最终也是参与回归模型的正常点最低要求，本文回归模型的选用强调突出局部数据特性，因此阈值的设定较小，为5左右。

同理，进入数据窗口的正常数据也仅取用邻近的部分参与回归模型，如果参与回归的数据太多，风速范围大，拟合精度会有所降低，尤其在数据密度低时，问题会进一步放大。图 是扩大窗口范围和参与回归计算的数据个数，在相同的不良数据识别结果下与默认参数进行效果对比。其中窗口大小扩大为24，最多选用参与重构的邻近点数量扩大到30，追求拟合的稳定性。

从结果来看，使用了更多正常点参与回归模型可能引入了大量噪声，功率特性曲线上的重构点更加紧凑的同时，整体数据分布更加分散。综合来看，小窗口运行下的重构结果更加优良，且运行速度更快。



1. **结论与展望**

本文针对风电场运行数据中存在的异常点问题，提出并实现了一套基于分箱LOESS局部加权回归与物理规则约束相结合的异常检测与窗口化的数据重构方法。在异常数据检测部分，对风功率数据按风速分箱局部回归拟合，设置动态残差阈值，同时结合风电机组实际运行要求划定物理规则，从数据测试的结果来看，本体提出的识别方法能够有效识别多种类型的异常数据，有利于提升风电场数据的准确性和可信度；在已有异常识别结果的数据重构部分，对检测出的不良数据采用邻域线性回归插值和物理约束修正，保留了数据的连续性和局部特征，通过数据搜索的防止避免了异常点对后续分析的干扰，测试重构后的功率特性曲线平滑合理，对风电场理想发电性能评估具有积极意义，另外，对不良数据的识别结果也为风电场大数据的健康管理与智能分析提供了一定参考价值。

当然，由于个人水平有限，提出的方法还存在许多不足，在第三章和第四章的测试结果分析中已有提到。从研究拓展角度来看，本文选用的LOESS回归方法不适用于高维的数据分析，未来的大研究方向对预测模型的精度要求会进一步提高，许多方法已经开始引入SCADA中的多维参数如温度、湿度、电机转速包括风机运行工况等来提高检测的全面和准确性。另外，未来的不良数据研究会聚焦于时序特征与深度学习方法，提高对隐蔽性异常的识别能力，再结合专家知识和机组状态信息，对异常点进行更加详细的划分，以实现针对性修正。

风电场数据异常检测与重构是提升风电智能运维和大数据价值的重要环节。随着技术进步与实际需求的提升，该领域将持续发展，并为新能源行业的数字化转型和智能化管理作出贡献。