

基于人工智能的风电出力功率预测

结 题 报 告

指导老师： 陈金富

负 责 人： 曲嘉骏 U202112266

项目成员： 董启翰 U202112254

项目成员： 王乐天 U202112269

项目成员： 李振宇 U202112543

项目成员： 李东朔 U202112540

2024. 4. 26

目 录

一、项目目标	4
(一) 项目背景	4
(1) 风力发电的前景	4
(2) 风电功率预测的重要性	4
(3) 人工智能在风电功率预测中的应用	4
(二) 项目目标与要求	5
(1) 寻找影响风电功率因素	5
(2) 建立高精度风电功率预测模型	6
二、项目实施方案	6
(一) 数据收集和数据预处理	6
(1) 数据集的收集	6
(2) 数据预处理	7
(二) 模型搭建	9
(1) 数据分解	9
(2) 模型建构	10
(3) 多模型叠加	14
三、项目研究成果	14
(一) 模型预测效果可视化	14
(二) 预测结果误差指标	16
(三) 预测结果分析与总结	16
四、项目创新点	17

（一）综合考虑多维因素	17
（二）数据分解	17
（三）综合应用多种人工智能技术	17
五、人员分工	17
六、经费使用情况	18
七、参考文献	18

一、项目目标

(一) 项目背景

(1) 风力发电的前景

随着对清洁能源需求的增加和技术的进步，风力发电已成为全球最主要的可再生能源之一。它具有无污染、可再生、资源丰富等优势，能够降低碳排放并推动可持续发展。随着技术的成熟和成本的下降，预计未来风力发电将继续增长，成为全球能源结构的重要组成部分。

(2) 风电功率预测的重要性

风电功率预测的意义在于帮助电网管理者和风电厂商合理规划和运营发电系统。准确的功率预测能够提前预知风电场产生的电能，帮助平衡电网负荷，避免电力供需不平衡，降低系统成本，提高电力利用率，促进可再生能源的大规模应用和可持续发展。

(3) 人工智能在风电功率预测中的应用

人工智能算法具有自动化、智能化、数据量大的优点，能够在超大规模的数据量中快速找出其映射关系，从而实现统计预测。通过机器学习算法建立风电发电量短期风电功率与影响因素之间的内在联系，将该联系训练成模型。通过调用训练好的模型，便可以预测短期风电功率。

目前常用的人工智能风电功率预测法主要是支持向量机法（SVM）和神经网络法。人工神经网络法在处理非线性问题上优势明显，自适应、自学习能力强，因此被广泛应用于风电功率预测上。有研究建立了以数值天气预报为基础，建立基于人工神经网络的风电功率预测系统，实现了与能量管理系统的无缝连接。也有文献提出一种基于数值天气预报以及人工神经网络的混合型风能预测系统，该系统能够依据特定风电机组或风场的发电量的历史数据对输出数据进行修正。为进一步提高预测精度，国内学者在预测模型和算法等方面进行了深入的理论研究。利用粗糙集理论分析影响风速预测的主要因素，将其作为中长期风速预测模型的附加输入，建立粗糙集混沌神经网络预测模型^[1]。

最新理论研究中，相关研究计算了区域和单个风电场的风电功率预测的误差，分析发现误差的减小取决于区域的大小，利用空间平滑效应预测区域风电功率提供了一种减小误差的新思路。也有研究将一种时间自适应的quantile-

copula方法应用于风电功率概率预测，并且讨论了如何选取对不同变量建模的核函数。另外有研究首先将风速序列进行小波分解，并对各分解信号采用自适应小波神经网络进行回归预测，再通过前馈神经网络将风速预测值转化风电功率预测值。文献[2]将风机数据和数值天气预报数据转化为风速向量，通过多个观察点预测风速和风向，再将风速预测值通过功率曲线转换成风电功率预测值。

（二）项目目标与要求

（1）寻找影响风电功率因素

影响风力发电功率的因素包括风速、风向、气温、空气密度、海拔高度、地形地貌、风机状态和设计参数等。风速是最主要的因素之一，风速越高，风机产生的功率就越大。某容量为2000kW的风电机组为例，风电机组的功率转换公式为：

$$G = \begin{cases} 0 & F < F_{in} \\ \frac{A\pi W^2 \rho v^3}{2} & F_{in} < F \leq F_r \\ G_K & F_r < F \leq F_{out} \\ 0 & F > F_{out} \end{cases}$$

式中：F为真实风速； F_{in} 为风电机组的切入速度； F_{out} 为风电机组的切出速度；A为风电机组的能量转化因子；W为风扇的叶片直径； ρ 为气体浓度； G_K 为风电机组的额定出力。通过该式可知，在不同风速下，风电机组的输出功率不同，大致可以分为以下四种情况：

- 1、在实际风速未达到风机的切入风速时，风机不转，产生的风电功率为0。
- 2、在实际风速超过切入风速后，风电机组的出力会随着风速的增大而增大，与风速的三次方成正比。
- 3、因风电机组的发电功率有限，当风速达到额定风速时，机组的输出功率不会随着风速的增长而继续增长，而是保持运行在额定输出功率状态。
- 4、在实际风速等于或超过切出风速的情况下，风电机组的出力会被切断。
风力过高，会造成风电机组输出功率过高，会切断与电力系统的联系。

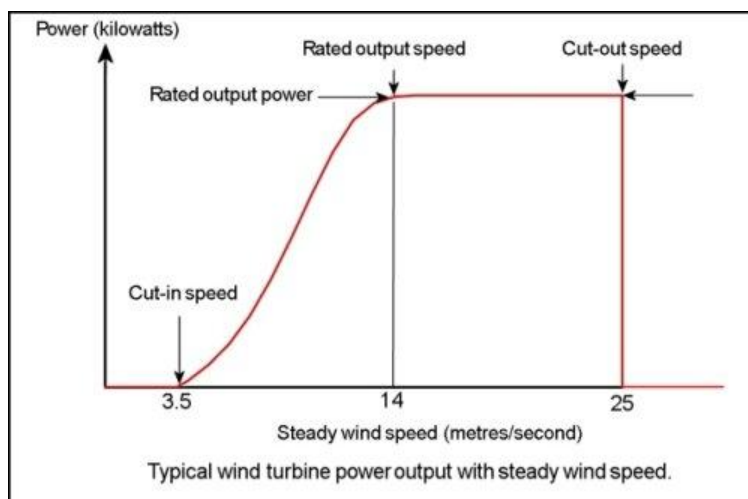


图1 理论风电功率-风速曲线图

风向也是影响风电输出功率的重要因素，由于机组的偏航调整装置存在一定的滞后性，风电机组可能在相同风速下有不同的输出功率。前排的风电机组吸收风能后，风速下降，后排的机组所产生的功率必然会减小，导致整体的风电场功率下降，通常称这种现象为尾流效应。此外，风机的状态、运行方式以及设计参数如桨距角度、叶片长度、旋转速度等也会对功率输出产生影响。综合考虑这些因素我们最终选择选取瞬时风速、30秒平均风速、10分钟平均风速、风向、温度、桨距角度等六个重要因素，作为本项目的六个参数。

(2) 建立高精度风电功率预测模型

由于风速和负荷的时空波动性，风电发电的不确定性给电网调度和负荷预测带来了挑战，因此，建立高效的风电-负荷场景生成与削减模型对于提高准确性至关重要。本项目需要收集大量历史风速数据和风力发电机组性能特征，利用多种人工智能技术，并考虑多种如风速、风向、气温等因素，旨在构建一个高准确性的风电功率预测模型。提高预测精度并尽可能降低误差，以实现对未来功率输出的准确预测，帮助风电厂优化运营计划、提高发电效率。

二、项目实施方案

(一) 数据收集和数据预处理

(1) 数据集的收集

数据集的采集是项目实施的关键步骤，直接影响后续模型建立和预测准确性。本团队经过线上与线下的长时间、多渠道收集，最终得到两个可用的数据集：

- 1、上海大石头风电场的数据集。该数据集包括瞬时风速、30秒平均风速、10分钟平均风速、风向、温度、桨距角度和功率等7个维度的108万个数据，时间颗粒度较高（30s记录一个数据），但时间跨度和连续性较差，用于相关性分析
- 2、某海上风力发电平台数据集。包括风速、风向、空气湿度等影响因素，具有良好的时间连续性，但数据量较少，包含16444个数据。

(2) 数据预处理

数据预处理包括异常值处理、缺失值处理和相关性检验三个方面。首先我们进行数据集的异常值处理。根据规定的运行规则，我们筛选了数据集中的异常值。这些规则包括：

- 1、保留风机正常运行状态的数据。
- 2、清除功率小于零的数据。
- 3、清除未达到切入风速时功率大于零的数据。
- 4、清除大于额定功率1.2倍的数据。

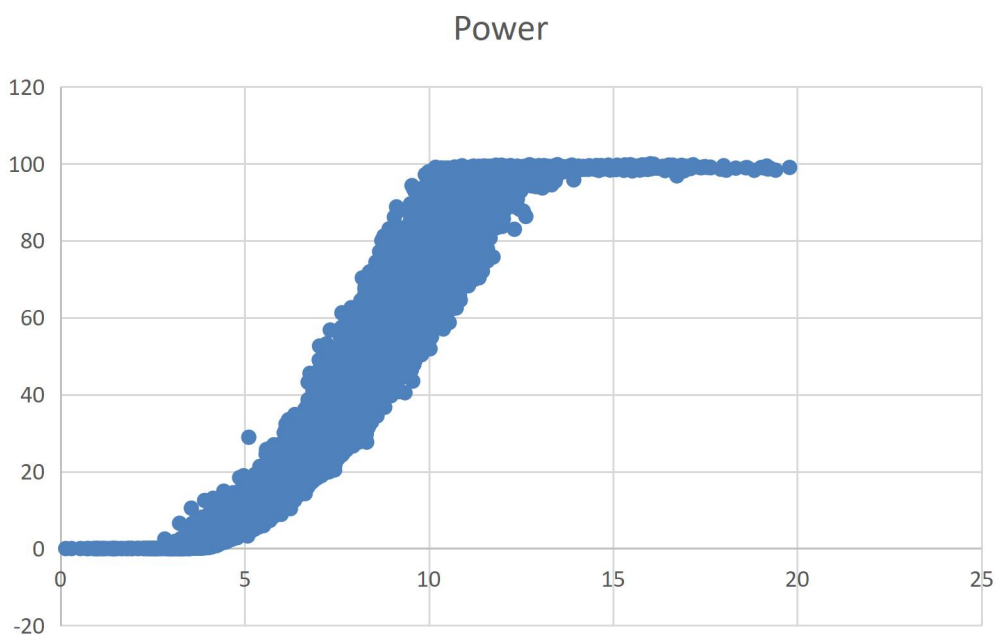


图2 根据风机规则清除异常值后的风机功率风速散点图

接着我们采用了基于密度的聚类算法DBSCAN来排除异常值。DBSCAN (Density-Based Spatial Clustering of Applications with Noise) 是一种常用的密度聚类算法，用于将数据点划分为具有高密度的区域（簇）和低密度

区域（噪声）。

算法原理：人工选择适当的邻域半径 ϵ 和最小邻域数据点数MinPts的值。对于数据集中的每个点，计算其 ϵ -邻域内的点数。如果这个数大于等于MinPts，则该点被视为核心点。连接处于彼此 ϵ -邻域内的核心点，形成一个或多个簇。将未分配到任何簇的非核心点标记为噪声或边界点。边界点位于某个簇的 ϵ -邻域内，但是不是核心点。

这样，DBSCAN通过探索数据点的密度分布，将具有足够高密度的点划分为簇，并识别出低密度区域中的噪声点。该算法的优点之一是它能够处理不同形状和大小的簇，并且不需要事先指定簇的数量。

经过多次调试，我们确定模型参数 $\epsilon=0.5$ ，MinPts=300。

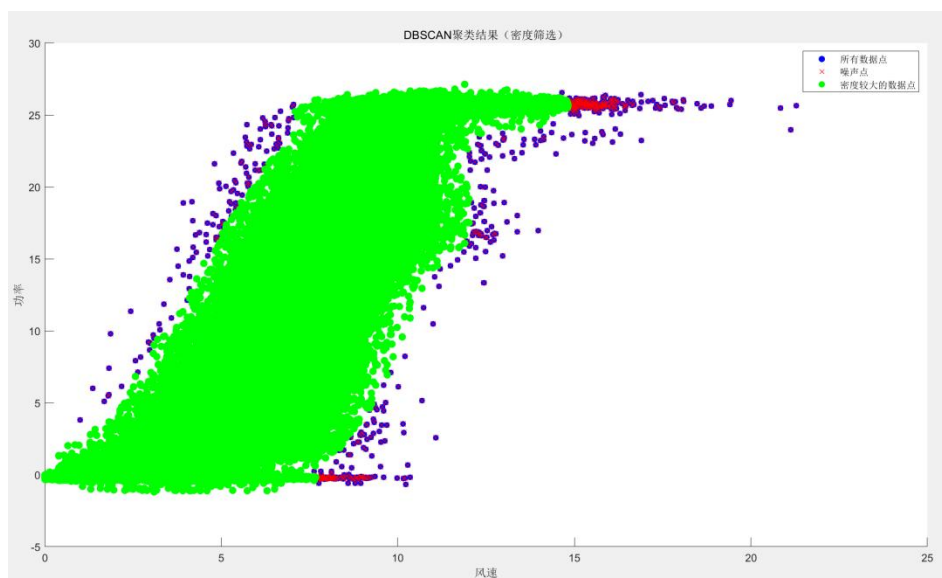


图3 根据DBSCAN算法清除异常值后的风机功率风速散点图

多种因素可能导致风电机组的异常，导致数据集中的数据缺失，有以下处理方法：删除法、插值法、均值/中位数/众数填充、模型预测填充。鉴于数据空缺位置不太多，我们选择了插值法处理数据缺失，数据量增加五分之一。

经过插值处理，数据表现出更好的连续性，对于后续建模和预测具有重要意义，能够提高数据质量、保持数据连续性、避免信息损失、提高模型稳定性，并增加建模样本量，从而为模型建立和预测工作提供更可靠的数据基础。

最后，我们通过计算相关系数矩阵并绘制热力图进行相关性检验，通过计算相关系数矩阵并绘制热力图，直观反映风电功率与各个因素的相关性。

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

在图4热力图中，不同颜色的方块代表相关系数的大小，深色表示相关性较强，浅色表示相关性较弱。通过观察热力图，可以直观地了解风电功率与其他因素之间的相关性，以指导后续建模和分析工作。我们发现，在数据集中，风电功率与风速呈强相关性，与气温和桨距角有一定程度的相关性，而与风向和内部温度相关性较弱。

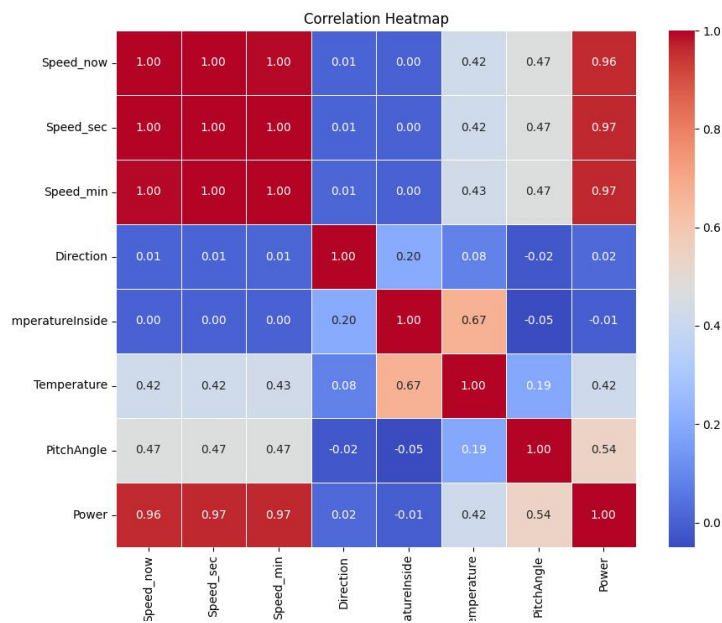


图4 相关性检验热力图

(二) 模型搭建

(1) 数据分解

在风电功率数据中，因受季节变化、气象条件等影响，其波动性较大且具有随机性，直接将原始数据输入模型难以捕捉到这种复杂的规律。季节性分解是一种把时间序列分解为趋势项、季节项和余项的过滤过程。通过这种分解，我们可以将数据拆解成三个主要部分：趋势、季节性和残差。趋势部分反映了数据的长期变化趋势，例如随着时间的推移，风电功率可能呈现出逐渐增长或减少的趋势。季节性部分则显示了数据在不同季节内的周期性变化，例如春季和夏季可能有较高的风电功率，而冬季则可能较低。而残差则是指无法通过趋势和季节性解释的随机波动，可能受到其他未知因素的影响。

对趋势和残差进行建模可以更准确地预测未来的数据趋势，帮助我们了解数据的长期变化趋势并做出相应的决策。而季节性部分则提供了对数据在不同季节内波动规律的深入了解，有助于我们制定更加精准的运营策略和预测方案，例如在高风电功率季节采取更多的发电策略，以提高发电效率。

另外，将数据分解后，可以针对不同的部分使用不同的预测模型。例如，可以针对趋势部分采用线性模型，针对残差部分采用深度学习模型等。不同模型对不同数据进行预测，各自产生的误差可能会相互抵消，从而降低总体的预测误差。这种叠加效应有助于提高整体预测的准确性，并更好地利用数据的信息。

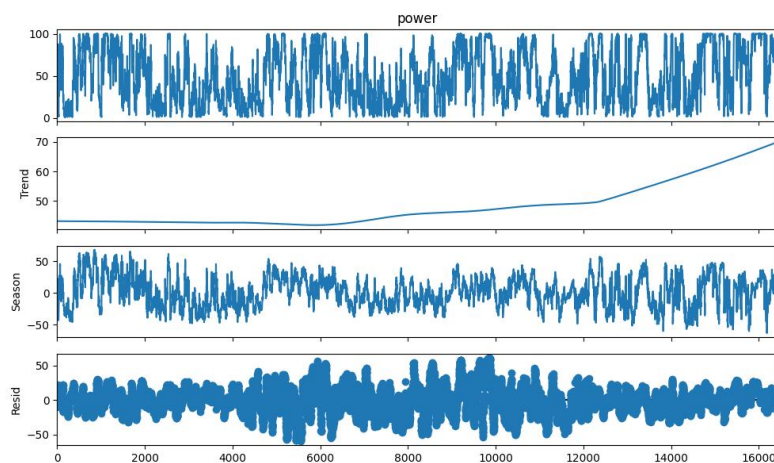


图5 季节性数据分解

(2) 模型建构

在模型建立阶段，我们综合考虑了风电功率数据的特点，根据数据分解的结果，采用了多元非线性回归、ARIMA-X和LSTM模型来分别处理趋势、季节性和残差。针对趋势部分的建模，考虑到风电功率在时间变化上有着简单的非线性趋势，我们选择了多元非线性回归模型。多元非线性回归预测模型是通过样本数据进行非线性函数拟合，再采用拟合函数来达到预测响应值的目的，常用的函数形式有幂函数、指数函数、对数函数等。我们广泛尝试各种函数对趋势数据进行拟合，最终得到了较为精确的二次函数拟合结果。表达式为：

$$y = 1.791 \cdot 10^{-7} x^2 - 0.0016x + 45.16$$

该模型拟合得到的R2系数为0.973，可以较好地反映趋势数据的非线性变化趋势。SARIMA-X模型用于处理季节性部分。传统的ARIMA模型是将非平稳时间序

列先经过d阶差分平稳化，再对得到的平稳时间序列利用自回归(AR(p) process)和滑动平均过程(MA(q) process)，并通过样本自相关系数(ACF)和偏自相关系数(PCF)等数据对建立的模型进行辨识，同时还提出了一整套的建模、估计、检验和控制方法。设 $\{y_t\}$ 为零均值的平稳时间序列，p 阶自回归 q 阶滑动平均的ARMA (p, q) 公式表述为：

$$Y_t - \phi_1 Y_{t-1} - \dots - \phi_p Y_{t-p} = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q}$$

ARIMA (p, d, q) 模型中的d是差分阶数，风电功率的时间序列数据一般都是非平稳的，差分是平稳化的途径之一，差分后的ARIMA建模过程基本与ARMA相同。如果时间序列不具稳定性，一般先采用ADF检验，选取差分阶数后将 y_t 变平稳，再结合赤池信息准则AIC和贝叶斯信息准则BIC对模型进行最佳参数选择^[3]。对比ACF和PACF图，我们确定ARIMA模型中p, q, d参数分别为1, 0, 1。

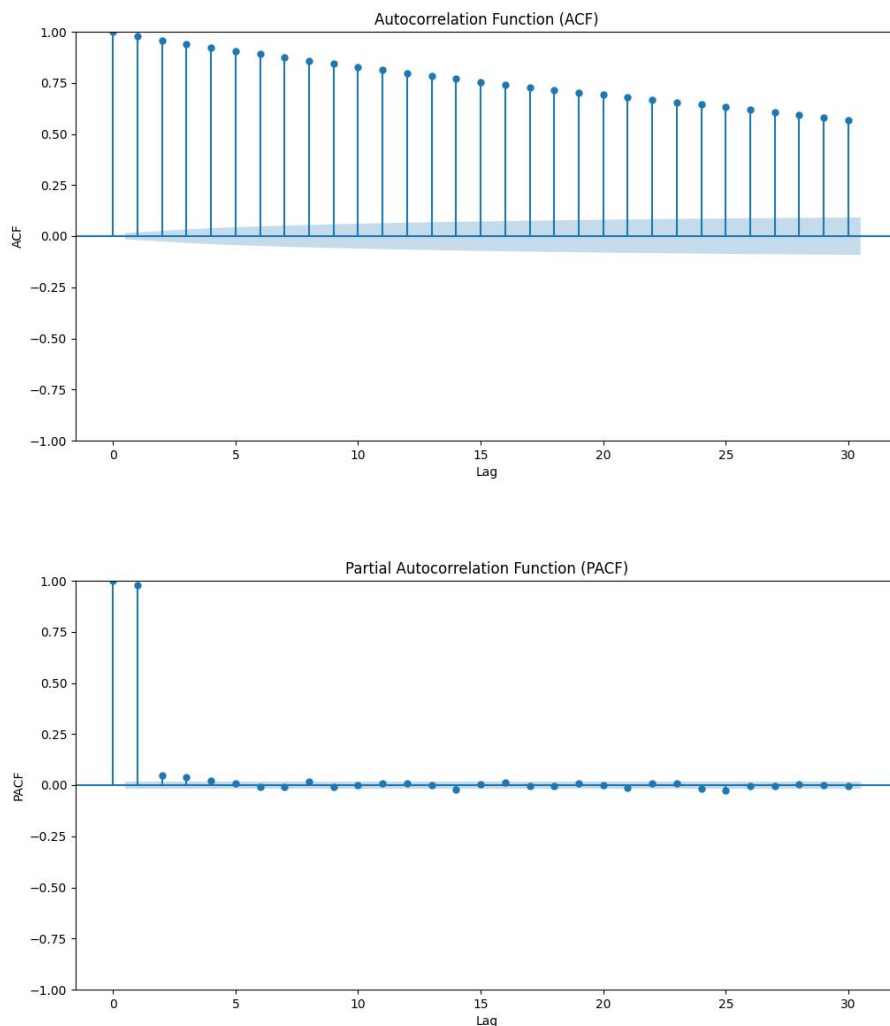


图6 样本自相关系数(ACF)和偏自相关系数(PCF)分析结果

而对于本项目的时序数据，我们进一步扩展模型为能捕捉季节特征的SARIMA模型。SARIMA (Seasonal ARIMA) 可以支持带有季节性成分特点的时间序列数据。它在ARIMA模型的基础上增加了4个季节性参数，分别是3个超参数(P, D, Q)和季节性周期参数s。其中，超参数P, Q, D, S分别代表季节自回归的阶数，季节移动平均的阶数，季节差分次数和季节/周期长度。考虑本项目使用数据为以月份为周期，我们确定以上参数为1, 0, 1, 12。

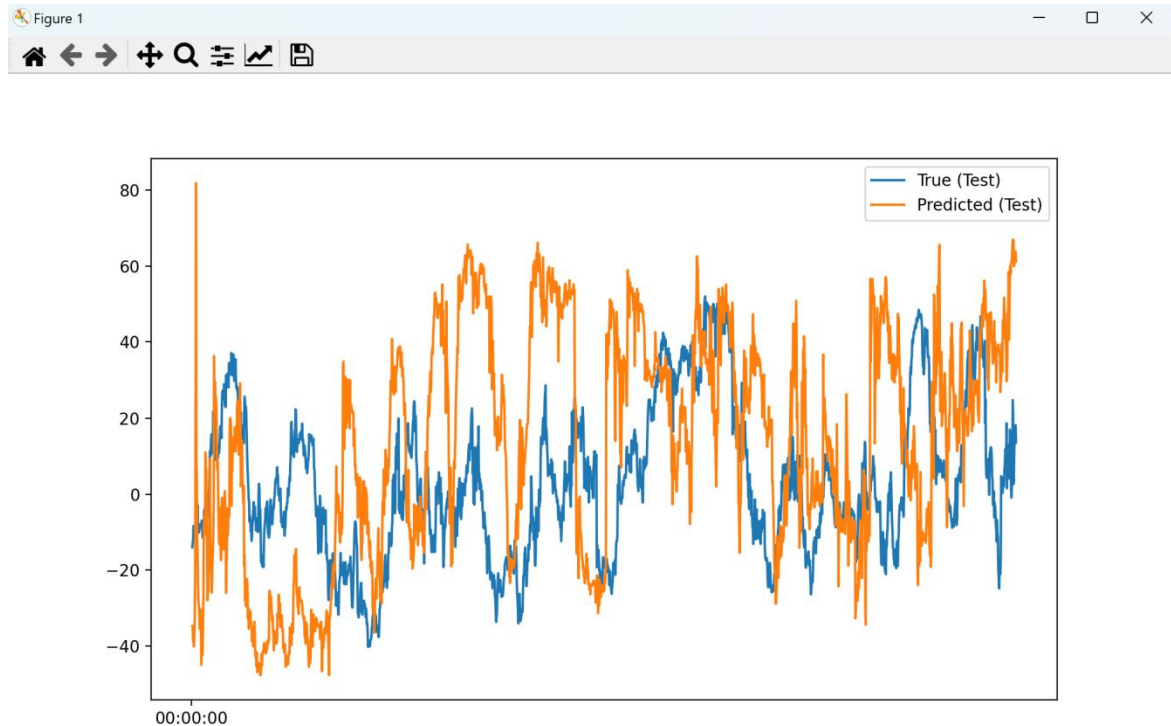


图7 SARIMA单因素模型预测效果

尽管经过了多次调参和调试，只有一维数据输入的SARIMA模型预测效果仍然较差。因此考虑到风电功率与风速等因素具有显著相关性，我们引入更复杂的SARIMA-X模型进行进一步的尝试。相比于传统ARIMA模型，ARIMA-X模型引入外部因素以提高模型对时间序列数据特征的捕捉能力。该模型能更精确处理风电功率数据复杂波动性的情况，提高预测准确率。能显著提高预测准确率，特别适用于具有大波动性的时间序列数据。使用该模型我们得到了不错的预测效果。

考虑到风电功率数据的复杂时序特征，我们选择了LSTM模型来建模残差部分。LSTM是一种适用于序列数据的深度学习模型，由于风电功率数据具有较强的时序相关性和复杂的波动性，传统的模型可能无法很好地捕捉其内在规律，

而LSTM模型能够更好地处理这种情况，提高预测的准确性和稳定性。

LSTM模型的关键是引入了一组记忆单元，使得网络能够学习何时遗忘历史信息以及何时用新信息更新记忆单元。在 t 时刻，记忆单元 C_t 记录了到当前时刻结束的所有历史信息并受3个“门”控制：遗忘门 f_t 、输入门 i_t 和输出门 o_t 。3个门的元素值在 $[0, 1]$ 之间。遗忘门 f_t 控制每一个内存单元需要多少信息，输入门 i_t 控制每一个内存单元加入多少新的信息，输出门 o_t 控制每一个内存单元输出多少信息^[4]。

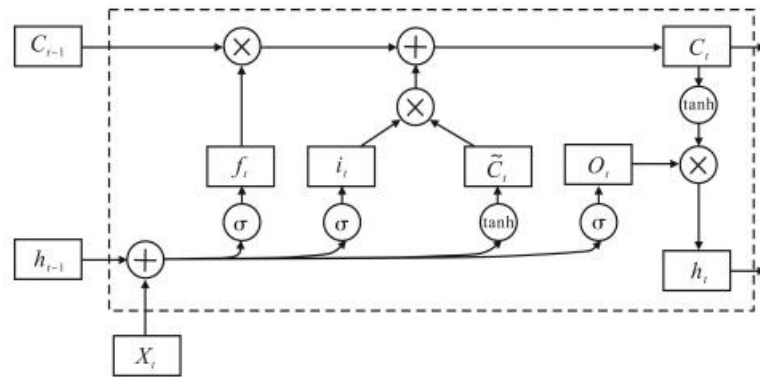


图8 LSTM模型计算结构

在构建用LSTM预测残差模型的过程中，我们首先对预处理后的残差数据进行了归一化处理，以便于模型的训练和收敛。然后，将数据划分为训练集和测试集，其中训练集用于模型的训练，测试集用于评估模型的预测效果。在构建LSTM模型时，输入层包括时间戳、风速、风向等多个维度的数据，以实现预测的准确性。

在具体参数选择上，需要考虑以下几点：窗口大小(window_size)决定模型每次观察的时间步长数，需平衡模型复杂性和捕获时间序列模式的能力；批量大小(batch_size)表示每次更新模型权重时所用样本数量，需根据计算资源和模型性能确定；迭代次数(epochs)决定模型在训练集上的训练次数，需要平衡拟合程度和过拟合风险；隐藏层神经元数量(hidden_dim)应考虑模型复杂度和数据特征；训练集比例(train_ratio)确定训练集在总数据集中的比例，需保证模型充分学习数据模式且具备泛化能力。

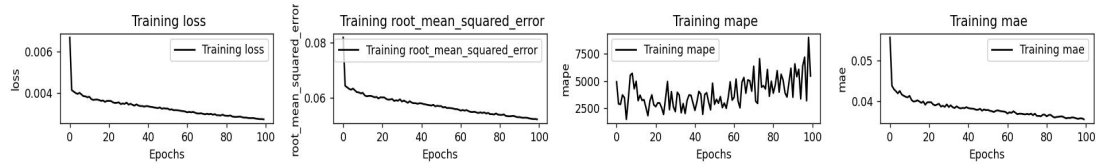


图9 未经过调参的误差曲线

基于训练结果和经验，我们选择 $\text{window_size} = 64$ $\text{batch_size} = 32$ $\text{epochs} = 100$ $\text{hidden_dim} = [32, 16]$ $\text{train_ratio} = 0.8$ 。此外，我们引入了回调函数以防止训练的过拟合现象产生。通过对训练集的迭代训练和LSTM神经网络具有的门控机制进行处理，LSTM模型可以学习到数据中的复杂时序规律，并对未来的残差进行有效预测。

(3) 多模型叠加

我们将以上三种模型经测试集训练后的结果进行相加，得到最终的预测结果，并与实际的测试集进行比较，根据最终的残差和误差指标评价模型的预测效果。

三、项目研究成果

(一) 模型预测效果可视化

利用python中的matplotlib库，我们对各个模型的预测效果进行了可视化绘图，以直观展示模型的预测效果。图10，11，12，13分别呈现了多元非线性回归拟合趋势数据，S-ARIMAX模型预测季节性数据，LSTM模型预测残差数据以及最终模型叠加的预测效果。可以看出，各个模型中预测值和实际的测试集数据值趋势基本相同，误差较小，说明均实现了较为准确的预测。

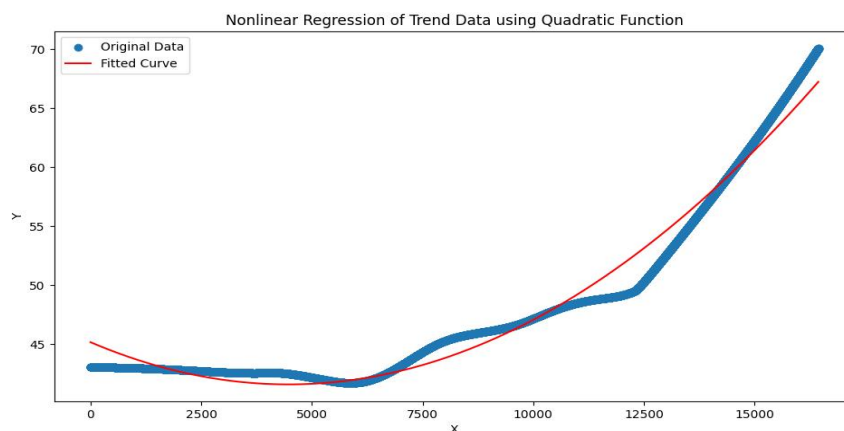


图10 多元非线性回归拟合趋势数据

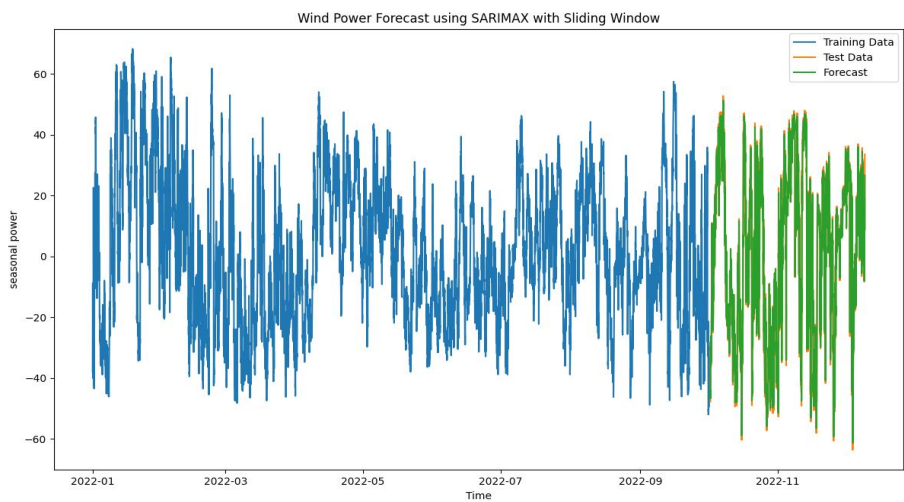


图11 SARIMA-X模型预测季节性数据

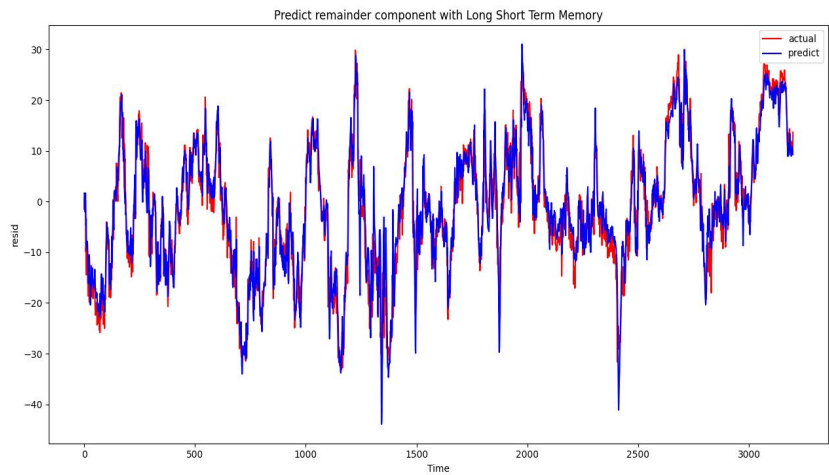


图12 LSTM模型预测残差数据

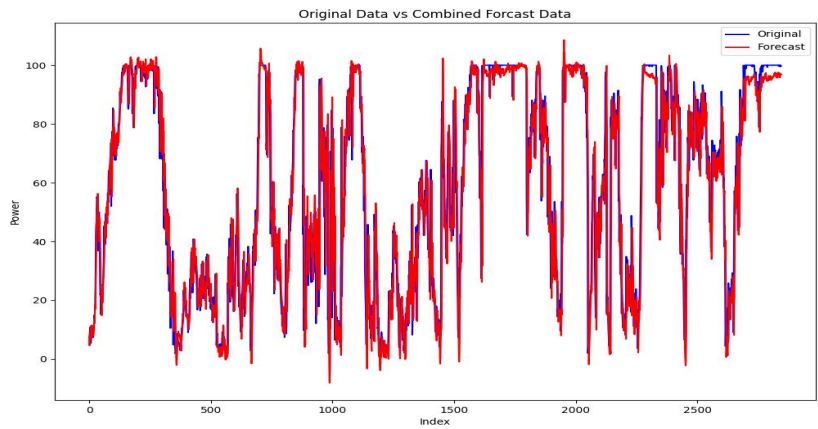


图13 最终叠加预测完整数据

(二) 预测结果误差指标

风力发电具有不确定性和模糊性，给功率预测带来了极大的困难，导致预测结果与真实值之间存在偏差，甚至在预测过程中，可能会发生各种失误，导致预测结果不理想，如数据收集过程中的操作失误、数据收集设备的老化损坏等，这些因素导致的偏差一般用具体数值来评估，即误差。

在LSTM预测模型中，我们采用的模型评估指标主要包括平均绝对百分比误差(Mean Absolute Percentage Error, MAPE)、均方根误差(Root Mean Square Error, RMSE)、均方误差(Mean squared error, MSE)及决定系数(R-Square)。MAPE、RMSE、MSE的值越小以及R2的值越接近1，表示风电功率预测效果越好。

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right|$$

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y}_i - y_i)^2}$$

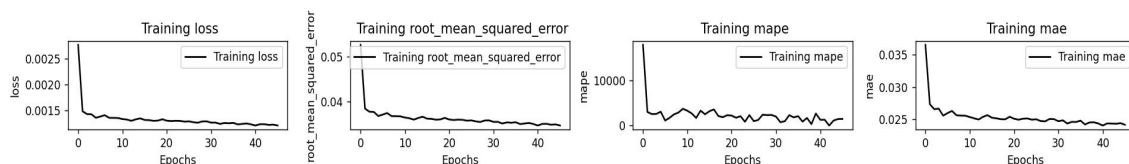


图14 LSTM模型预测结果的评估指标

各个模型的评价指标结果如下：

表1 评价指标结果表

	多元非线性回归	SARIMA-X	LSTM	叠加模型
R2	0.973	0.868	0.914	0.946
MSE		152.939	13.346	39.394
RMSE		12.366	3.777	6.276
MAPE		29.980	-0.099	89.192

(三) 预测结果分析与总结

由以上评价指标可知，各个模型经过调参后均取得了较为准确的预测结果。这表明我们选择的模型可以正确捕捉对应的数据特征。经过叠加之后的模型相较于单独的模型具有更好的预测效果， R^2 值更接近于1，各个误差评价指标都有了一定程度的减小。通过数据分解和多模型叠加的方式本项目在风电功率预测精度上取得了不错的结果。

四、项目创新点

（一）综合考虑多维因素

在建立预测模型时，项目充分考虑了影响风力发电功率的多种因素，包括风速、风向、气温等。通过综合考虑这些因素，构建了全面的预测模型，使模型能够更准确地捕捉到风力发电系统的运行规律，从而提高了预测的准确性和稳定性。虽然这样多维数据的输入增加了数据收集和模型构建的难度，增加了模型运行的时间，但也大大提高了预测的精度。相比于传统的单时间输入的ARIMA模型，多维输入更可以反映风电功率的复杂变化规律。

（二）数据分解

风力发电具有随机性，波动性的特点，因而风电功率数据在时间上呈现的规律复杂，起伏大，难以捕捉随时间的变化规律。项目采用季节性分解的方法，将风力发电功率数据拆分为趋势、季节性和残差三个部分。针对每个部分的特点，选择了不同类型的模型进行建模。这种数据分解的策略使得不同模型能够更好地适应数据的特征，提高了预测的精度和可靠性。

（三）综合应用多种人工智能技术

本项目不局限于单一的预测模型，而是综合运用了时间序列分析方法、深度学习神经网络模型、以及多元非线性回归模型对于分解得到的不同数据进行针对性的分析建模。通过结合这些不同类型的模型，可以充分挖掘风力发电数据中的各种特征信息，从而提高预测模型的准确性和稳定性。

五、人员分工

曲嘉骏：项目负责人。负责整个团队工作的协调统筹与整体规划的制定与落实。

负责数据预处理、ARIMA-X模型代码的编写与调试、多模型叠加整合，项目结题的答辩。

董启翰：负责数据预处理、LSTM模型代码的编写、调试与测试，项目结题报告的撰写。

王乐天：负责数据预处理、多元非线性回归模型代码的编写与调试，项目结题答辩PPT的制作。

李振宇：负责进行风电场数据的收集，风电场数据的整合，项目结题答辩PPT的制作。

李东朔：负责进行风电场数据的收集与整理，风电场数据的整合，项目结题报告的撰写。

陈金富导师：全程对项目团队进行指导与资源提供，在每一不同阶段提出相关建议和规划，并对团队提出的问题和需要的帮助进行及时的回应与解答，协助团队学生完成相关任务。

六、经费使用情况

未使用经费

名称	单价（元）	数量	总计（元）

七、参考文献

- [1] 叶林，赵永宁. 基于空间相关性的风电功率预测研究综述[J]. 电力系统自动化, 2014, 38(14):126-135.
- [2] 薛禹胜，郁琛，赵俊华, 等. 关于短期及超短期风电功率预测的评述[J]. 电力系统自动化, 2015, 39(6):141-151.
- [3] 吴玉霞，温 欣. 基于ARIMA模型的短期股票价格预测. 南京财经大学. 2016
- [4] 王英林，陶 蕾，付佳宇. 基于改进 LSTM 的海上风电功率短期预测方法. 中国电力工程顾问集团华北电力设计院有限公司. 2023
- [5] 周志华. 《机器学习》. 清华大学出版社. 2016