

文章编号:0254-0096(2019)12-3594-11

# 基于聚类与非参数核密度估计的风电功率预测 误差分析

张晓英<sup>1~3</sup>, 张晓敏<sup>1~3</sup>, 廖顺<sup>4</sup>, 陈伟<sup>1~3</sup>, 王晓兰<sup>1~3</sup>

(1. 兰州理工大学电气工程与信息工程学院, 兰州 730050; 2. 甘肃省工业过程先进控制重点实验室, 兰州 730050;  
3. 兰州理工大学电气与控制工程国家级实验教学示范中心, 兰州 730050; 4. 国网四川省电力公司凉山供电公司, 西昌 615000)

**摘要:** 针对实际风电场风电功率预测误差呈现出季节、时序和功率变化特性, 提出基于聚类分析与非参数核密度估计的方法研究风电功率预测误差的概率分布特性。采用聚类分析的方法进行月份、时段的缩减, 有效地将误差特性相似的误差数据归为一组, 从而在考虑误差分布多样性的基础上兼顾误差分布的整体趋势。在此基础上, 考虑功率特性, 采用非参数核密度估计方法进行风电功率预测误差概率分布拟合并采取基于迭代的窗宽求解方法。通过拟合精度评价指标验证所提方法的适用性及有效性。

**关键词:** 风电功率预测误差; 分段聚类; 非参数估计; 窗宽求解; 拟合精度

**中图分类号:** TM614

**文献标识码:** A

## 0 引言

自然存在的风能是发展前景较好的绿色环保能源之一, 大规模开发建设风电场利用风能进行发电缓解能源匮乏已成为全球能源资源利用的大趋势。风能固有的随机性以及风能预测水平的限制使得风电功率预测误差成为难以消弭的存在。随着全球风电的迅猛发展, 风电的不确定出力给系统调峰调频及备用容量的确定带来的影响不容忽视<sup>[1]</sup>。在现有风电功率预测水平条件下深入研究风电功率预测误差, 以求在最大程度上抑制风能高度随机性给电网运行带来的危害显得尤其重要。

国内外针对风电功率预测误差分布特性已经做了一系列的研究。目前, 相关研究采用的概率分布拟合方法主要分 2 大类: 参数法和非参数法。近年来, Copula 理论也逐渐应用于新能源不确定性的研究<sup>[2,3]</sup>。参数估计方法以假设误差数据符合现有的某个具体的概率分布为前提, 再尝试利用这种分布进行数据拟合。早期应用于预测误差研究的主要是正态分布<sup>[4]</sup>, 但后续的研究发现预测误差并不

严格服从正态分布, 表现出一定的有偏性及尾部拖延性。文献[5,6]采用峰度可变的 Beta 分布来描述具有拖尾特性的风电功率预测误差。文献[7]指出 Beta 分布在某些预测功率区段上出现概率密度无穷大的情况。文献[8]在经典正态分布上添加 3 个系数对模型标准差进行调整, 可灵活描述类正态分布, 但该方法参数的确定比较复杂且只考虑尖峰性, 忽略对有偏性的描述。文献[9]提出一种带三参数的通用分布模型并将其用于预测误差的拟合, 但在预测误差分布集中的功率段, 对尖峰和多峰的拟合效果较差。文献[10]将带双参数的  $t$  分布用于拟合。针对单一概率分布对预测误差多峰拟合精度较差的缺陷, 文献[11]提出两段指数分布模型进行双峰预测误差的概率分布拟合, 该模型的两段形状参数独立, 可较好地拟合双峰误差, 但在多峰值情况下的应用有局限性。文献[12]利用偏正态分布对有偏性较好的描述能力, 将混合偏正态分布模型应用于预测误差的描述中, 该模型在描述有偏性的同时能较好地兼顾尾部厚重特性, 但该模型参数较多, 求取过程较复杂。文献[13]通过混合高斯分

收稿日期: 2017-06-14

基金项目: 国家自然科学基金(51867015; 51767017); 甘肃省基础研究创新群体项目(18JR3RA133); 甘肃省高校协同创新团队项目(2018C-09)

通信作者: 张晓英(1973—), 女, 硕士、教授, 主要从事电力系统分析与控制方面的研究。245659219@qq.com

布对比分析不同预测方法下的拟合效果。参数法中预先假设的概率分布对于数据的适应性决定这种方法下概率模型的拟合精度,对于多峰值的拟合效果较差。

非参数核密度估计是非参数估计法中应用较广的方法之一。该方法无需事先假设预测误差服从某个具体的标准参数分布即可对其直接进行拟合。文献[14,15]使用非参数核密度估计对预测误差进行建模分析,但文献[14]未考虑误差的季节特性,文献[15]在对预测误差的功率分段问题上考虑太过单一,且没有考虑时序特性。

综上所述,目前对风电功率预测误差的研究,主要存在2个问题,一是多数确定性的模型不能很好地兼顾描述风电功率预测误差呈现的多峰、尖峰和厚尾特性;二是在分析风电功率预测误差时,对于季节特征、时序特性和功率特性的考虑不够周全。其中对多峰、尖峰部分的拟合也是预测误差分析的难点所在。本文在上述研究的基础上,通过分析风电功率预测误差的季节、时序和功率特性,首先采用聚类方法对预测误差的分段情况进行优化,其次利用非参数核密度估计对分段处理后的预测误差进行拟合,最后通过与传统经验分布模型的拟合评价指标对比验证所提方法的适用性及有效性。

## 1 风电功率预测误差特性分析

### 1.1 预测误差数据预处理

根据国家能源局颁布的“风电场功率预测预报管理暂行办法”<sup>[16]</sup>要求参与并网的风电场向调度机构上报次日24小时每15分钟共计96个时间点的风电数据,本研究所用数据采集以15分钟为时间分辨率,预测方法为数值天气预报法。

为避免在某些时间段风速很低的情况下,相对误差失去运行指导意义,本文所分析的风电功率预测误差选用绝对误差来表征。 $t$ 时段绝对误差可用式(1)表示:

$$e_t = \hat{P}_t - P_t \quad (1)$$

式中, $\hat{P}_t$ ——风电功率预测值; $P_t$ ——风电功率实际值。

为后续计算更加方便,将绝对误差数据以风电场额定装机容量 $P_{\text{cap}}$ 为标准进行归一化处理,使得预测值和实际值取值范围为 $[0, 1]$ 之间,相应的误

差取值在 $[-1, 1]$ 之间。归一化可用式(2)表示:

$$e = \frac{\hat{P}_t - P_t}{P_{\text{cap}}} \quad (2)$$

### 1.2 误差分布特性统计

对预测误差分布特性进行拟合分析的问题上,不可片面地认为未来某一时刻的风电功率预测误差服从全年的整体分布。已有研究表明,随着风电功率预测值的增大,相应的预测误差也会随之表现增加的趋势。除此之外还应考虑风速、风电的实际出力在各个季节、各个时段的特征差异。图1为某个风电场全年的整体预测误差概率分布直方图,图2为其1月份的误差概率分布直方图。由图1和图2可看出1月份的预测误差分布情况和全年12个月的整体分布差异较明显,整体分布呈现出尖峰形式的类正态分布,而落实到具体的月份上预测误差则表现出有偏性。图3则以统计形式给出各月份预测误差的绝对平均误差,表明各月份间预测误差的差异。

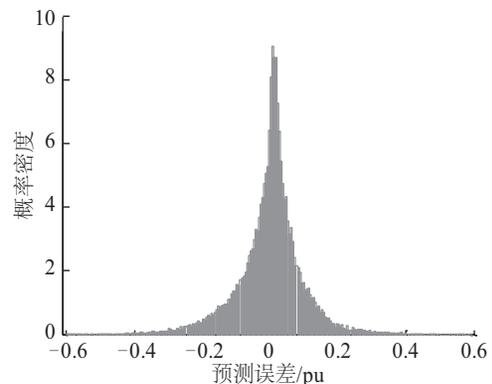


图1 风电功率全年整体预测误差

Fig. 1 Full-year general prediction error of wind power

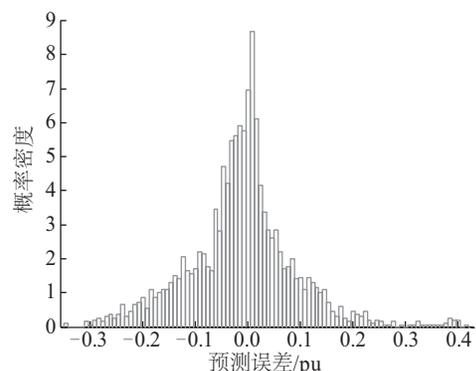


图2 1月风功率预测误差

Fig. 2 Prediction error of wind power in January

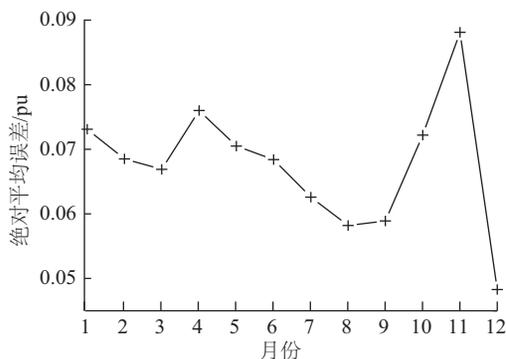


图3 12个月份的绝对平均误差

Fig. 3 Absolute mean error of 12 months

统计全年纵向时间序列的预测误差数据,即以一年为统计范围,分别统计对应于24个时段的数据。计算分析24个时段的绝对平均误差和预测功率平均值,可得图4所示的走势变化图。从图4可看出,在全年统计范围下,每天24个时段的风能预测功率的绝对平均误差存在一定的差异,且其走势变化大致上跟随相应时段预测功率的平均值而变化,即预测功率平均值增大,该误差也随之增加。

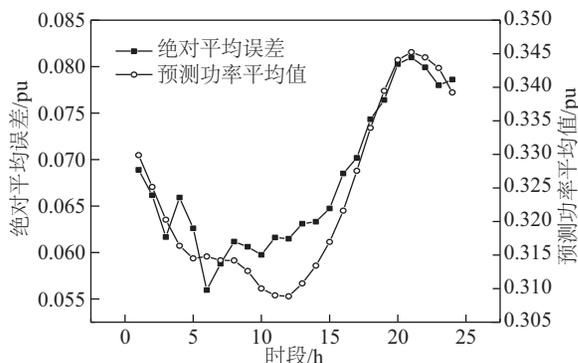


图4 24个时段的预测误差

Fig. 4 Prediction error of 24 time intervals

从整体上把握预测误差随风电功率预测值的变化情况对于分析误差的实际分布具有重要的指导意义。图5给出全年的风电的预测功率与预测误差一一对应的散点图。从图5可清楚地看到在预测功率较小的情况下,预测误差的分布较为集中。因此,必须根据预测功率值的大小来分段分析预测误差的分布。

## 2 基于聚类分析的分段处理

### 2.1 功率段的划分

由1.2节分析可知,深入分析预测误差的实际

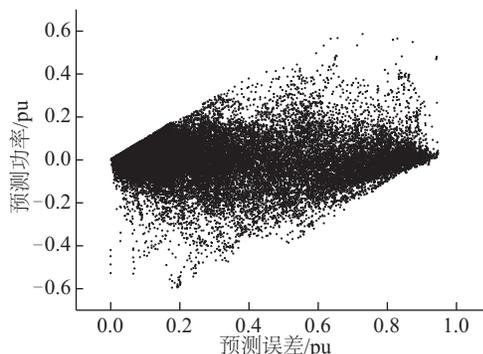


图5 预测功率-预测误差散点图

Fig. 5 Predictive power-prediction error scatter plot

分布情况需要考虑季节、纵向时序和预测功率值。合理地进行分段能够最大程度上体现预测误差分布的多样性。

考虑12个月份、24个时段以及预测功率在 $[0,1]$ 范围粗略划分为10个区间的分段方式得到的分段数较多,对误差的描述过于细化且预测误差在10个区间上的数据量分布差异太大,文中采取分段缩减方法以保证在尽可能合理考虑误差分布多样性的基础上兼顾整体趋势,既能考虑到误差的区间特性又可从全局把握误差的分布情况。

对于预测误差在预测功率层面上的分段问题,文献[17]根据预测误差散点图的分布规律,粗略地将其划分为3段。本文在此基础上首先将预测误差在 $[0,1]$ 上均匀划分为10个区间,区间跨度为0.1。统计结果显示在预测功率小于0.4的4个区段上,其数据总体占有率为66.7%,其中预测功率在 $[0,0.1]$ 的区段数据量占有率达28%,而 $[0.9,1]$ 的区段仅占0.72%。由此可见,这种粗略分段导致的数据量在各区段的失衡是很严重的,不利于体现预测误差分布规律。借鉴文献[14]对功率段数缩减的方法,本文将预测功率先根据数据量占有率均分为5段,区段数缩减一半,每个区间内数据量为 $b = a/5$ , $a$ 为数据样本总量,根据划分的区间范围,可作适当调整保证每个区间的数量在 $b$ 的2%左右波动。由此划定预测功率的分段,即 $[0, 0.065]$ ,  $[0.065, 0.165]$ ,  $[0.165, 0.33]$ ,  $[0.33, 0.63]$ ,  $[0.63, 1.0]$ ,这样的划分方法得到的各个区间段内数据量基本持平。

### 2.2 采用聚类分析进行区段缩减

将预测误差数据在月份上划分为12组、时序

上划分为 24 组会导致数据分组数过多,且风速在各月份和时序上也呈现一定的变化规律。聚类分析算法的目标是将数据对象自动归入到有意义的类别中,使得类内相似性尽可能高,类间差异性尽可能大。本文采用聚类分析方法将预测误差数据特性相似的月份(时段)划分在一起,归为一类。

在聚类分群中,分群指标是反映数据特征的量,是聚类划分的根据,单一分群指标会导致分类结果不准确<sup>[18]</sup>。对于 12 个月份和 24 个纵向时段的缩减划分,本文从表征误差特征的指标中选取绝对平均误差、误差平均值和均方根误差 3 项作为聚类分群的指标,将模糊 C-均值聚类算法(fuzzy C-means algorithm, FCM)<sup>[19]</sup>用于各月份、时刻的聚类分析。FCM 算法的基本思想是将需要聚类的对象记为  $X=[x_1, x_2, \dots, x_k, \dots, x_n]^T$ ,其中  $x_k$  为需进行聚类的对象之一,它有  $m$  个表征特性的指标。如将  $X$  分为  $c$  类,初始化每一类的聚类中心,在聚类进行的过程中通过计算各对象与各聚类中心的距离来确定分类,这是一个迭代循环的过程,直到满足迭代终止条件得到最终的聚类结果。此处对月份、时刻的聚类应分开进行,以 12 个月份的聚类为例,具体步骤简述如下:

1) 记  $X$  为 12 个月份的 3 项指标构成的样本数据矩阵,即  $X=[x_{11}, x_{12}, \dots, x_{k1}, \dots, x_{kn}]^T$ 。其中  $x_k$  为第  $k$  个样本,即第  $k$  个月份的 3 项指标构成的向量  $x_k=[x_{k1}, x_{k2}, \dots, x_{km}]$ ,因为预测误差采用标么值,此处的  $X$  为归一化的  $n \times m$  矩阵, $n$  为需聚类的样本总数, $m$  为聚类指标项数,此处  $n=12, m=3$ 。

2) 设定聚类类别数  $c$  ( $2 \leq c \leq n$ ) 和迭代停止阈值  $\varepsilon$ ,随机初始化聚类中心  $V=\{v_1, v_2, \dots, v_p, \dots, v_c\}$ ,其中  $v_p$  为第  $p$  类的聚类中心,  $v_p=[v_{p1}, v_{p2}, \dots, v_{pm}]$  和隶属度矩阵  $U=(u_{pk})_{c \times n}$ ,其中,  $u_{pk}$  表示第  $k$  个样本  $x_k$  属于第  $p$  类的隶属度值。

3) 计算分群指标构成的样本数据矩阵  $X$  中各样本  $x_k$  到各初始聚类中心的欧式距离  $d_{pk}$ 。

$$d_{pk} = \sqrt{(x_{k1} - v_{p1})^2 + \dots + (x_{km} - v_{pm})^2} \quad (3)$$

4) 隶属度  $u_{pk}$  的计算,  $0 \leq u_{pk} \leq 1, \sum_{p=1}^c u_{pk} = 1$ 。

$$u_{pk} = 1 / \sum_{l=1}^c \left( \frac{d_{pk}}{d_{lk}} \right)^2 \quad (4)$$

5) 更新聚类中心  $v_p$ 。

$$v_p = \frac{\sum_{k=1}^n (u_{pk})^2 x_k}{\sum_{k=1}^n (u_{pk})^2} \quad (5)$$

6) 设  $iter$  为循环迭代次数,按下式计算目标函数。

$$F = \sum_{p=1}^c \sum_{k=1}^n (u_{pk})^2 (d_{pk})^2 \quad (6)$$

7) FCM 聚类算法的目标函数的值是随着迭代次数的增加而下降的。判断迭代终止条件是否满足,即  $F(iter) < \varepsilon$  算法结束,比较数据样本  $x_k$  归属于各类别的隶属度值,将样本标记为隶属度最大的那一类。统计各类别的样本得到分类结果;否则算法转向步骤 3),继续进行迭代计算。

### 3 预测误差的非参数核密度估计

假设  $e_1, e_2, \dots, e_N$  为风电功率预测误差  $e$  在分段处理后的含  $N$  个样本的一组数据。 $f(e)$  为预测误差的原始概率密度函数,则  $\hat{f}(e)$  为  $f(e)$  的核密度估计,表示如下:

$$\hat{f}(e) = \frac{1}{Nh} \sum_{i=1}^N K(h^{-1}(e - e_i)) \quad (7)$$

式中,  $N$  —— 一组数据的样本数量;  $h$  —— 窗宽,也称光滑参数;  $K(u)$  —— 核函数,  $u = h^{-1}(e - e_i)$ ;  $e_i$  —— 预测误差数据中的第  $i$  个样本值。

在式(7)中影响估计结果的因素有 2 点:一是核函数  $K(\cdot)$ ,二是窗宽  $h$ 。常见的核函数有均匀(Uniform)核、高斯(Gaussian)核、Epanechnikov 核、四次(Quartic)核。文献[20]指出核函数的不同选择在核密度估计中不敏感,当  $N$  足够大时,它对估计结果的影响不大。

本文选用较常用的高斯核函数,即  $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$ ,则估计  $\hat{f}(e)$  用式(8)表示:

$$\hat{f}(e) = \frac{1}{Nh} \sum_{i=1}^N \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} \left( \frac{e - e_i}{h} \right)^2\right] \quad (8)$$

核密度估计精度的一种度量是均方误差(mean squared error, MSE),  $MSE(\hat{f}(e)) = E[\hat{f}(e) - f(e)]^2 =$

$\text{var}(\hat{f}(e)) + [\text{bias}(\hat{f}(e))]^2$ , 其中  $\text{var}(\hat{f}(e))$  表示估计的方差,  $\text{bias}(\hat{f}(e))$  表示估计的偏差。窗宽  $h$  对  $f(e)$  起着局部光滑的作用。在样本给定的情况下, 估计性能的好坏主要取决于窗宽  $h$  的选择是否得当。过小的  $h$  会使偏差得到改善, 但估计的随机性影响增加, 导致  $\hat{f}(e)$  呈现出不规则的形状。而过大的  $h$ ,  $\hat{f}(e)$  过度平滑, 使  $f(e)$  的细节特征不能显示出来。因此, 使用非参数核密度估计拟合时, 窗宽的选择最为关键。

窗宽的选择采用的方法一般有正态参考规则 (normal reference rule, NR)<sup>[21]</sup>, 最小二乘交叉验证法 (least squares cross validation, LSCV)<sup>[22]</sup>。正态参考规则是一种通过经验公式直接确定窗宽的方法, 即有  $h_{opt} = 1.06\hat{\sigma}n^{-1/5}$ , 其中  $\hat{\sigma}$  可取  $\min\{S, Q/1.34\}$ ,  $S$  为样本标准差,  $Q$  为样本  $e_1, e_2, \dots, e_N$  的 75% 与 25% 的分位数之差。而最小二乘交叉验证法基于最小化积分平方误差 (integrated squared error, ISE) 的一种自动选择带宽的方法。本文选择文献[23]提出的一种迭代求取窗宽的方法确定核密度估计的窗宽, 该方法是对基于正态参考规则确定窗宽方法的改进。

首先根据正态参考规则得到窗宽  $h_0 = 1.06\hat{\sigma}N^{-1/5}$ , 将  $h_0$  代入式(8)得到一个核密度估计为:

$$\hat{f}(e) = \frac{1}{Nh_0} \sum_{i=1}^N \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{e-e_i}{h_0}\right)^2\right] \quad (9)$$

将式(9)中的  $\hat{f}(e)$  当作真实的密度, 同时考虑渐进积分均方误差 (asymptotic MSE),  $AMISE[\hat{f}(e)] = \int MSE[\hat{f}(e)]de$  求得一个窗宽  $h_1$ , 将  $h_1$  赋值给  $h_2$ , 将  $h_2$  代入式(8)得到另一个核密度估计式将其作为真实概率密度, 重复上述步骤得到一个趋于稳定的窗宽即为最优窗宽  $h_{opt}$ 。

## 4 算例分析

为验证上述所提分段考虑方法及基于迭代法求取窗宽的非参数核密度估计模型的有效性和适用性, 本文采用比利时电力运营商 Elia 公开的风电场 2016 年 1~12 月的风电功率预测值及实际生产值数据, 数据采集时间分辨率为 15 min。对这些数

据采用上面提到的方法进行处理。

### 4.1 数据划分

采用 2.2 节方法对 12 个月、24 个纵向时段进行聚类划分时, 首先考虑 12 个月份的季节特性将聚类分类数  $c$  取为 4, 聚类结果将 11 月单独归为一类, 于是考虑将  $c$  设定为 3; 在对时序进行划分时按照上午、下午、晚上、夜间的时间界定将  $c$  值取为 4。具体分类结果如表 1、表 2 所示。

表 1 月份划分结果

类别	月份
第 1 类	1、3、4、10、11
第 2 类	2、5、6
第 3 类	7、8、9、12

表 2 时段划分结果

类别	时段
第 1 类	1、2、4、5、11、12、13、14、15
第 2 类	3、6、7、8、9、10
第 3 类	19、20、21、22、23、24
第 4 类	16、17、18

### 4.2 拟合精度评价指标

在对预测误差数据进行划分的基础上进行概率拟合, 采用拟合精度评价指标对各拟合模型的拟合误差进行对比分析。然而, 在现有的文献中衡量评价拟合误差的指标众多, 对此在学术界还没有统一的规定。在文献[24]中, Hansen 和 Lunde 建议尽可能采用多种不同的指标作为拟合模型精度的评判标准。基于对这种方法的思考, 现采用以下指标作为分布模型拟合精度评价标准。

#### 1) 效率系数和一致性系数

效率系数和一致性系数常用于水文领域的模型精度性能评价<sup>[25]</sup>。效率系数是对模型拟合值与实际分布值的平均值之间的相似程度的描述, 而一致性系数则反映模型拟合结果和实际分布之间的一致性程度。这 2 种评价指标均属于横向误差指标, 从序列形态的角度出发描述模型峰值在横向坐标上的拟合情况, 指标值越大表明分布模型对原始分布的解释能力越强, 对数据的拟合效果也更好。

原始预测误差数据的概率分布坐标为  $(x_i, y_i)$  ,  $i = 1, 2, \dots, N$ ,  $N$  为序列长度;拟合分布模型的坐标为  $(x_i, \hat{y}_i)$  ,  $i = 1, 2, \dots, N$ , 模型拟合误差  $E_i = \hat{y}_i - y_i$  。精度评价指标的计算是在原始分布和拟合分布的横坐标  $x_i$  是一一对应的条件下,计算纵坐标的一系列变化的过程。

效率系数定义如式(10)所示:

$$E = 1.0 - \frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{\sum_{i=1}^N (\bar{y} - y_i)^2} \tag{10}$$

式中,  $\hat{y}_i$  ——分布模型拟合值;  $y_i$  ——误差概率密度实际值;  $\bar{y}$  ——实际值的平均值。理论上  $E$  的取值范围为  $-\infty \sim 1$  ,  $E$  的值越接近 1 表明分布模型拟合值与误差概率密度实际值越接近。

分析式(10)可知,效率系数  $E$  是 1 减去分布模型的残差平方和与总离差平方和比率得到的差值,因此,式(10)也可用式(11)表示:

$$E = 1.0 - \frac{SSE}{SST} \tag{11}$$

式中,  $SSE = \sum_{i=1}^N (\hat{y}_i - y_i)^2 = \sum_{i=1}^N E_i^2$  为分布模型的残差平方和(sum of squares for error);  $SST = \sum_{i=1}^N (\bar{y} - y_i)^2$  为分布模型总离差平方和(sum of squares for total)。

一致性系数定义为:

$$d = 1.0 - \frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{\sum_{i=1}^N (|\bar{y} - \hat{y}_i| + |\bar{y} - y_i|)^2} \tag{12}$$

$$= 1.0 - \frac{SSE}{\sum_{i=1}^N (|\bar{y} - \hat{y}_i| + |\bar{y} - y_i|)^2}$$

由式(12)可看出一致性系数的取值范围在  $[0, 1]$  之间,  $d$  值越接近 1 表明分布模型对原始预测误差的拟合性好,两者一致性越强。

2)绝对平均误差和均方根误差

绝对平均误差(mean absolute error,  $MAE$ )和均方根误差(root mean square error,  $RMSE$ )是 2 类衡量误差时使用最多纵向误差指标。  $MAE$  反映分布模型误差平均幅值的情况,而  $RMSE$  则说明模型误差的分散度,两者均越小越优,表明拟合分布模型在纵向上与原始概率密度越接近,即所用模型拟合精

度越高。

2 种指标具体定义如式(13)和式(14)所示:

$$MAE = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| = \frac{1}{N} \sum_{i=1}^N |E_i| \tag{13}$$

$$RMSE = \sqrt{SSE/N}$$

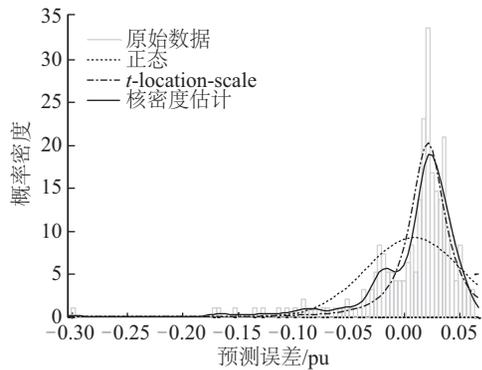
$$= \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} = \sqrt{\frac{1}{N} \sum_{i=1}^N E_i^2} \tag{14}$$

基于上述 4 种精度评价指标对比所选模型与传统经验分布模型的评价结果,分析所选模型的优越性及适用性。

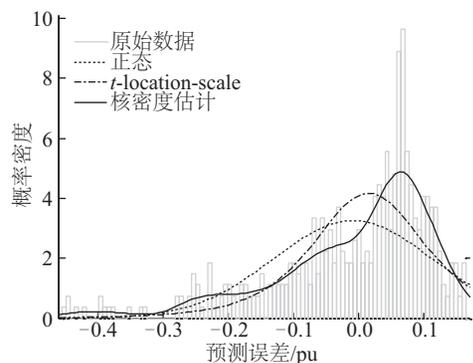
4.3 模型评价分析

4.3.1 不同拟合方法比较

基于 4.1 节的数据划分,预测误差数据按月份分 3 组,时段分 4 组,功率分 5 段进行。将月份、时段和功率段看作一个整体。如提取位于月份第 2 组、时段第 3 组和功率第 1 段的误差数据,用  $e\{2, 3, 1\}$  表示。下面就误差数据  $e\{2, 3, j\}$  , ( $j = 1, 2, 3, 5$ ),采用 2 种传统经验分布模型与非参数核密度估计对其进行拟合,并计算对应的精度评价指标。图 6 为各模型对应的分布拟合图,相应的评价指标见表 3。



a. 误差数据  $e\{2, 3, 1\}$



b. 误差数据  $e\{2, 3, 2\}$

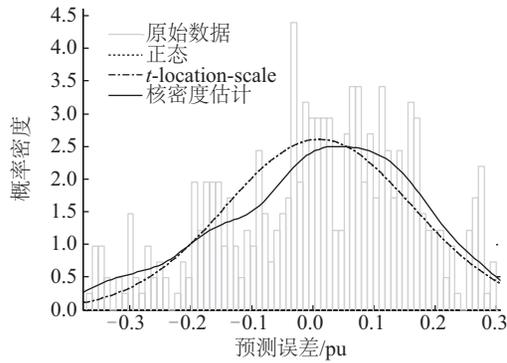
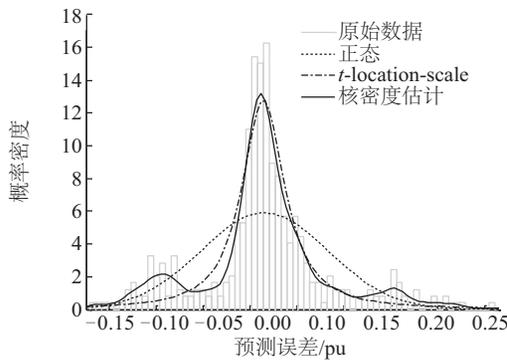
c. 误差数据  $e\{2, 3, 3\}$ d. 误差数据  $e\{2, 3, 5\}$ 

图6 各模型下的预测误差概率分布拟合

Fig. 6 Fitting of probabilistic distribution of prediction error under each model

从分布的实际形态分析:由图6可看出在月份、时段上处于同一分组里的误差数据在不同的功率段上分布差异很大,其中 $e\{2, 3, 1\}$ 和 $e\{2, 3, 5\}$ 的峰度和偏度分别为 $(15.0460, -2.8006)$ ,  $(4.3261, 0.2978)$ ,在概率分布上呈现出典型的多峰、尖峰和尾部拖延特性;而 $e\{2, 3, 2\}$ 和 $e\{2, 3, 3\}$ 的峰度和偏度分别为 $(4.5388, -1.2916)$ ,  $(2.5998, -0.4015)$ ,其概率分布虽然在峰值上表现为多峰,但峰值的变化波动很不规律。由于误差原始概率密度分布在形态上呈左偏或右偏,而具有无偏性的正态分布拟合在峰度上不够尖,高峰部分拟合能力不足,多峰拟合效果低,整体上正态分布拟合效果较差。具有尺度参数和位置参数的 $t$ 分布虽然较正态分布而言对多峰、尾部拖延的拟合效果较好,但对于多峰的拟合仍显得有些平滑,在峰值凸起部分有低估现象,在分布凹陷部分表现为高估实际分布。核密度估计法相对于其他2种分布模型能在拟合时较好地表现其多峰值部分。

表3 不同模型下各误差数据组拟合指标统计

Table 3 Fitting statistics of each error data set under different models

指标	模型	图6a	图6b	图6c	图6d
MAE	正态	1.7191	0.8198	0.5954	1.5153
	$t$ -location-scale	1.1700	0.7194	0.5954	0.9045
	核密度	0.8927	0.5196	0.5520	0.6313
RMSE	正态	3.8251	1.3158	0.7573	2.5649
	$t$ -location-scale	2.5349	1.2006	0.7573	1.3315
	核密度	2.1126	0.8499	0.6809	1.0084
E	正态	0.5512	0.4512	0.4516	0.4683
	$t$ -location-scale	0.7858	0.5438	0.4516	0.5566
	核密度	0.8512	0.7714	0.5566	0.9178
d	正态	0.7874	0.7918	0.8140	0.7667
	$t$ -location-scale	0.9381	0.8321	0.8140	0.9610
	核密度	0.9555	0.9111	0.8350	0.9165

从精度评价指标的角度分析:表3可清楚看出,在不同误差数据段上的MAE和RMSE均随正态分布、 $t$ -location-scale分布、核密度估计法呈现减小趋势;而效率系数 $E$ 和一致性系数 $d$ 则表现为依次增大趋势。这充分说明核密度估计法在指标的综合评价上优于其他2种模型,尤其是对 $e\{2, 3, 1\}$ 和 $e\{2, 3, 5\}$ 的误差分布,核密度估计相对正态分布和 $t$ -location-scale分布表现出的优势更明显。在此需要说明,在 $e\{2, 3, 3\}$ 的拟合上, $t$ -location-scale分布的形状参数 $v$ 为无穷大,该分布趋向于正态分布。

综上所述,对这类误差数据的概率密度,核密度估计法能较好地应用。因此,文中采用非参数核密度估计法进行风电功率预测误差的概率分布拟合。

#### 4.3.2 窗宽选取比较

为了提高非参数核密度估计法的拟合效果,本文采用上面提到的迭代法来求取窗宽 $h$ ,同时利用最初的正态参考原则(NR)及最小二乘交叉法(LSCV)生成窗宽,计算各窗宽求取方法下核密度估计的各拟合精度评价指标值。

这里仍以4.3.1节的误差数据为研究对象进行对比分析,3种窗宽求取法的拟合分布如图7所示,相应的指标变化及其窗宽参数分别由表4、表5给出。

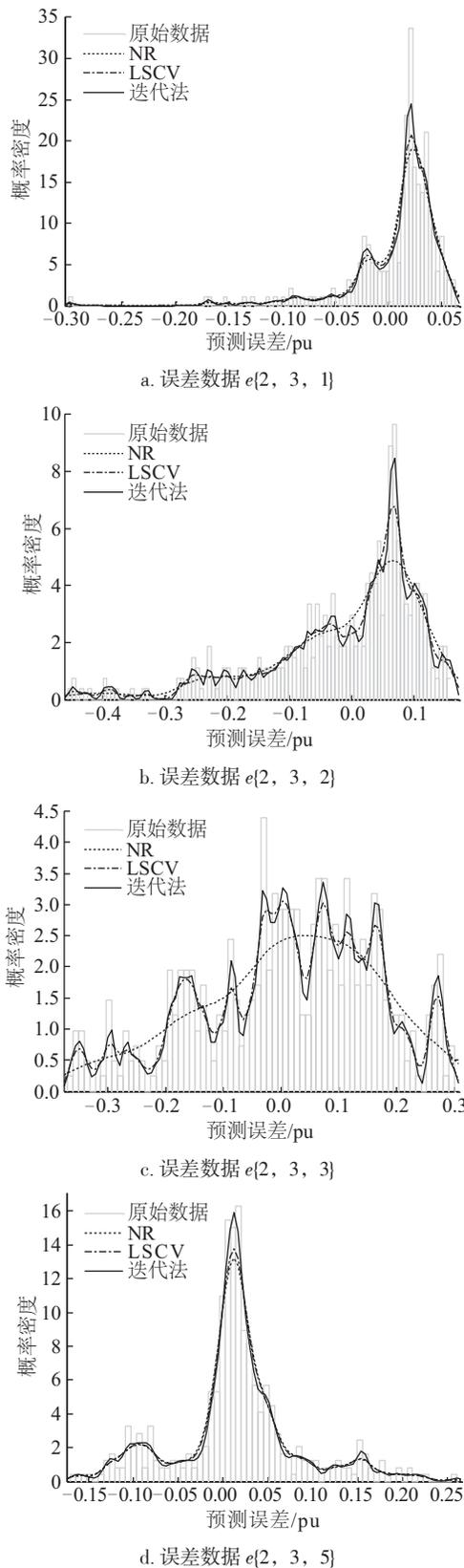


图7 3种窗宽选取方法下的误差概率分布拟合

Fig. 7 Fitting of probabilistic distribution of prediction error under three different ways of window width selection

表4 不同窗宽选取法下各误差数据组拟合指标统计

Table 4 Fitting statistics of each error data set under different ways of window width selection

指标	方法	图7a	图7b	图7c	图7d
MAE	NR	0.8927	0.5196	0.5520	0.6313
	LSCV	0.8357	0.4465	0.4262	0.6017
	迭代法	0.7373	0.3724	0.3696	0.5128
RMSE	NR	2.1126	0.8499	0.6809	1.0084
	LSCV	1.9489	0.6724	0.5243	0.9468
	迭代法	1.6305	0.5526	0.4503	0.8322
E	NR	0.8512	0.7714	0.5566	0.9178
	LSCV	0.8734	0.8569	0.7371	0.9276
	迭代法	0.9114	0.9033	0.9276	0.9439
d	NR	0.9555	0.9111	0.8350	0.9765
	LSCV	0.9633	0.9445	0.9149	0.9797
	迭代法	0.9759	0.9738	0.9433	0.9851

表5 窗宽参数值

Table 5 Window width parameter value

方法	图7a	图7b	图7c	图7d
NR	0.0079	0.0294	0.0489	0.0104
LSCV	0.0062	0.0102	0.0106	0.0092
迭代法	0.0038	0.0047	0.0065	0.0055

从图7可看出,对于非参数核密度估计,不同窗宽求取方法在误差概率密度的拟合上的区别主要表现在对尖峰的拟合上。采取迭代法在拟合上的优越性具体可由表4的精度评价指标的变化看出来,即对于几组误差数据的拟合,相比采用正态参考原则和最小二乘交叉法得到的窗宽进行拟合,本文所采取的迭代法的绝对平均误差MAE和均方根误差RMSE均较小;效率系数E和一致性系数d则较前2种方法更高。3种方法在不同指标上的变化趋势略有差异,如一致性系数d的变化,不同的误差数据随窗宽求取方法的不同变化趋势不一样,有的较平缓,有的变化较大,这与误差数据本身的分布特征有关,比如 $e(2, 3, 5)$ 这组数据在用正态分布经验法求取窗宽时,一致性系数d已较高,因此采取不同方法的变化较平缓。

通过以上对比分析可看出,非参数核密度估计方法对风电功率预测误差具有尖峰、多峰和尾部拖延特性的数据拟合能力较传统的经验分布模型

好。在此基础上,结合一种新的窗宽求解方法对几组误差数据进行拟合,不同方法下的各精度评价指标反映所用拟合方法的适用性及精确性。对德国某风电场 2016 年风电功率数据分析结果验证了所提方法对于其他风电场的实用性,由于篇幅所限,具体过程略去。

## 5 结 论

针对风电功率预测误差固有的尖峰、多峰和尾部拖延特性,分析预测误差在季节、时序和功率分段上的分布特性,在总结现有拟合方法的基础上,本文提出采用聚类分析的方法对误差数据进行分段,并采用非参数核密度估计进行概率密度拟合,得到结论如下:

1)各月份的风电功率预测误差与全年的误差分布特性差异较大,全年的误差分布在形态上呈尖峰形式的类正态分布,而各月份的分布大多呈多峰,尾部拖延特征。一天 24 个时段,不同功率段上的误差数据在误差分布上也不同。因此,研究风电功率预测误差时考虑季节、时序和功率可更好地刻画误差特征。

2)提出采用聚类分析的方法进行月份和时段的缩减,能有效地将误差特性相似的误差数据归为一组,避免过细描述预测误差而忽略整体特性。

3)在误差数据分组的情况下提出采用非参数核密度估计法进行概率密度拟合,且采取基于迭代的窗宽求解方法。各精度评价指标反映所用方法的优越性。

4)所提误差拟合方法可应用到后续的考虑风电不确定性的调度模式研究中。

### [参考文献]

- [1] 张丽英,叶廷路,辛耀中,等.大规模风电接入电网的相关问题及措施[J].中国电机工程学报,2010,30(25):1—9.
- [1] Zhang Liying, Ye Tinglu, Xin Yaozhong, et al. Problems and measures of power grid accommodating large scale wind power[J]. Proceedings of the CSEE, 2010, 30(25): 1—9.
- [2] Zhang Ning, Kang Chongqing, Xia Qing, et al. Modeling conditional forecast error for wind power in generation scheduling[J]. IEEE Transactions on Power Systems, 2014, 29(3): 1316—1324.
- [3] 赵唯嘉,张 宁,康重庆,等.光伏发电出力的条件预测误差概率分布估计方法[J].电力系统自动化,2015,39(16):8—15.
- [3] Zhao Weijia, Zhang Ning, Kang Chongqing, et al. A method of probabilistic distribution estimation of conditional forecast error for photovoltaic power generation[J]. Automation of Electric Power Systems, 2015, 39(16): 8—15.
- [4] Pinson P, Kariniotakis G N. Wind power forecasting using fuzzy neural networks enhanced with on-line prediction risk assessment [A]. Proceedings of IEEE Conference on Power Conference [C], Bologna, Italy, European Commission, 2003: 23—26.
- [5] Fabbri A, Gomze T, Roman S, et al. Assessment of the cost associated with wind generation prediction errors in a liberalized electricity market[J]. IEEE Trans on Power Systems, 2005, 20(3): 1440—1446.
- [6] Bludszuweit H, Dominguez-Navarro J A, Lombart A. statistical analysis of wind power forecast error[J]. IEEE Transactions on Power Systems, 2008, 23(3): 983—991.
- [7] Tewri S, Geyer C J, Mohan N. A statistical model for wind power forecast error and its application to the estimation of penalties in liberalized markets[J]. IEEE Trans on Power Systems, 2011, 26(4): 2031—2039.
- [8] 刘 斌,周京阳,周海明,等.一种改进的风电功率预测误差分布模型[J].华东电力,2012,40(2): 286—291.
- [8] Liu Bin, Zhou Jingyang, Zhou Haiming, et al. An improved model for wind power forecast error distribution [J]. Power System Technology, 2012, 40(2): 286—291.
- [9] Zhang Zhaosui, Sun Yuanzhang, Gao David Wenzhong, et al. A versatile probability distribution model for wind power forecast errors and its application in economic dispatch[J]. IEEE Transactions on Power Systems, 2013, 28(3): 3114—3125.
- [10] 刘立阳,吴军基,孟绍良.短期风电功率预测误差分布研究[J].电力系统保护与控制,2013,41(12): 65—70.
- [10] Liu Liyang, Wu Junji, Meng Shaoliang. Research on error distribution of short-term wind power prediction [J]. Power System Protection and Control, 2013, 41(12): 65—70.
- [11] 刘 芳,潘 毅,刘 辉,等.风电功率预测误差分段指数分布模型[J].电力系统自动化,2013,37

- (18): 14—19.
- [11] Liu Fang, Pan Yi, Liu Hui, et al. Piecewise exponential distribution model of wind power forecasting error[J]. *Automation of Electric Power Systems*, 2013, 37(18): 14—19.
- [12] 刘燕华, 李伟花, 刘冲, 等. 短期风电功率预测误差的混合偏态分布模型[J]. *中国电机工程学报*, 2015, 35(10): 2375—2382.
- [12] Liu Yanhua, Li Weihua, Liu Chong, et al. Mixed skew distribution model of short-term wind power prediction error[J]. *Proceedings of the CSEE*, 2015, 35(10): 2375—2382.
- [13] 杨茂, 董骏城. 基于混合高斯分布的风电功率实时预测误差分析[J]. *太阳能学报*, 2016, 37(6): 1594—1602.
- [13] Yang Mao, Dong Juncheng. Real-time prediction error analysis of wind power based on mixed Gaussian distribution model[J]. *Acta Energiæ Solaris Sinica*, 2016, 37(6): 1594—1602.
- [14] Wang Zhaoqing, Wang Chengfu, Liang Jun, et al. The fitting method of wing power forecast error under power-time dimension[A]. *The 5th International Conference on Electric Utility Deregulation and Restructuring and Power Technologies*[C], Changsha, 2015.
- [15] 刘立阳, 孟绍良, 吴军基. 基于风电预测误差区间的动态经济调度[J]. *电力自动化设备*, 2016, 36(9): 87—93.
- [15] Liu Liyang, Meng Shaoliang, Wu Junji. Dynamic economic dispatch based on wind power forecast error interval[J]. *Electric Power Automation Equipment*, 2016, 36(9): 87—93.
- [16] 国能新能[2011]177号. 风电场功率预测预报管理暂行办法[S]. 北京: 国家能源局, 2011.
- [16] Guoneng Xinneng [2011]177. The notification of wind power forecasting management interim measures[S]. Beijing: National Energy Board, 2011.
- [17] 赵书强, 王扬, 徐岩. 基于风电预测误差随机性的火储联合相关机会规划调度[J]. *中国电机工程学报*, 2014, (S1): 9—16.
- [17] Zhao Shuqiang, Wang Yang, Xu Yan. Dependent chance programming dispatching of integrated thermal power generation and energy storage system based on wind power forecasting error[J]. *Proceedings of the CSEE*, 2014, (S1): 9—16.
- [18] 夏世威, 白雪峰, 陈士麟, 等. 基于模糊聚类法的暂态稳定机组分群方法[J]. *电力系统自动化*, 2010, 34(2): 29—34.
- [18] Xia Shiwei, Bai Xuefeng, Chen Shilin, et al. A fuzzy method for clustering generators in transient stability analysis[J]. *Automation of Electric Power Systems*, 2010, 34(2): 29—34.
- [19] Bezdek J C. *Pattern recognition with fuzzy objective function algorithms*[M]. New York: Plenum Press, 1981.
- [20] Epanechnikov V A. Nonparametric estimation of a multidimensional probability density[J]. *Theory of Probability and its Application*, 1969, 14(1): 153—158.
- [21] Silverman B W. *Density estimation for statistics and data analysis*[M]. Boca Raton: CRC Press, 1986.
- [22] Bowman A W. An alternative method of cross-validation for the smoothing of density estimates[J]. *Biometrika*, 1984, 71(32): 353—360.
- [23] 胡蓓蓓, 宗刚. 非参数核密度估计在异方差模型中的应用[J]. *数量经济技术经济研究*, 2014, 10: 151—161.
- [23] Hu Beibei, Zong Gang. Application of Heteroscedastic models in non-parametric Kernel density estimation[J]. *The Journal of Quantitative & Technical Economics*, 2014, 10: 151—161.
- [24] Hansen P R, Lunde A. A forecast comparison of volatility models: Does anything beat a GARCH(1, 1)[J]. *Journal of Applied Econometrics*, 2005, 20(7): 873—899.
- [25] 王蕾, 王鹏新, 田苗, 等. 效率系数和一致性系数及其在干旱预测精度评价中的应用[J]. *干旱地区农业研究*, 2016, 34(1): 229—235.
- [25] Wang Lei, Wang Pengxin, Tian Miao, et al. Application of the coefficient of efficiency and index of agreement on accuracy assessment of drought forecasting models[J]. *Agricultural Research in the Arid Areas*, 2016, 34(1): 229—235.

## PREDICTION ERROR ANALYSIS OF WIND POWER BASED ON CLUSTERING AND NON-PARAMETRIC KERNEL DENSITY ESTIMATION

Zhang Xiaoying<sup>1-3</sup>, Zhang Xiaomin<sup>1-3</sup>, Liao Shun<sup>4</sup>, Chen Wei<sup>1-3</sup>, Wang Xiaolan<sup>1-3</sup>

(1. *College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China;*

2. *Key Laboratory of Gansu Advanced Control for Industrial Processes, Lanzhou University of Technology, Lanzhou 730050, China;*

3. *National Demonstration Center for Experimental Electrical and Control Engineering Education, Lanzhou University of Technology, Lanzhou 730050, China;* 4. *Liangshan Power Supply Corporation, State Grid Sichuan Electric Power Company, Xichang 615000, China*)

**Abstract:** According to the characteristics of season, time sequence and power change of the actual wind power prediction error, a new research method is proposed based on the clustering analysis and the non-parametric kernel density estimation to study the probability distribution of wind power prediction error. By using the method of the cluster analysis for months and time frame reduction, the data with the similar error characteristics is effectively classified to a group, which ensures the diversity and the overall trend of error distribution. On the basis of this, considering the power characteristics, using the proposed method to simulate the wind power forecast error probability distribution, the iterative method is adopted to solve the window width in the process. The applicability and effectiveness of the proposed method are verified by the evaluating indicator of fitting accuracy.

**Keywords:** wind power forecast error; segmental clustering; non-parametric estimation; solving of window width; fitting accuracy