基于最优箱宽直方图的牵引变电所负荷概率建模方法

车玉龙[1,2],王晓茹[2],吕晓琴[2],康永强[3]
(1. 兰州交通大学 自动化与电气工程学院,甘肃 兰州 730070;2. 西南交通大学 电气工程学院,四川 成都 611756;
3. 兰州交通大学 新能源与动力工程学院,甘肃 兰州 730070)

摘要:电气化铁路牵引负荷受众多随机因素的影响,较难用已知的参数估计模型来统一描述其概率模型。直方图是最简单直观且应用最广泛的一种非参数估计方法,但由其得到的概率密度形状极易受箱宽(或箱数)的影响。为得到牵引负荷概率分布的最佳直方图,提出一种基于最优箱宽建立其概率模型的直方图方法。由Poisson点事件构造直方图,通过平均积分平方误差定义价值函数,利用Sturges规则和平均移位直方图思想计算最优箱宽。采用3种不同采样间隔的牵引变电所负荷测试数据,通过概率分布形状和后验检验,与4种传统计算直方图箱宽的经验规则方法进行比较,验证了所提方法的有效性、适应性和稳健性。

关键词:牵引负荷;概率模型;直方图;价值函数;最优箱宽;平均移位
中图分类号:TM 922.3　　**文献标志码**:A　　DOI:10.16081/j.epae.202012006

0 引言

电气化铁路牵引变电所负荷是移动的大功率单相冲击性电力负荷,列车在运行过程中受到行车组织、线路条件和自然环境等因素的影响,具有很强的随机性和波动性。若能得到牵引负荷的概率密度函数(PDF),则能很容易得到其期望、方差和各阶矩等数字特征量。这对刻画牵引负荷的随机性和波动性,以及分析牵引负荷对电网的影响具有重要意义。

估计概率密度的方法可分为参数估计和非参数估计[1]。文献[2-5]分别采用正态分布、Beta分布、Lognormal分布和多项式函数等已知分布模型拟合了牵引负荷的概率模型。文献[6]基于大量实测数据,采用Origin软件通过分段曲线拟合方法对CRH2-200型动车组建立了谐波源概率模型。文献[7]基于实测数据利用模拟退火算法进行参数辨识,建立了牵引变电所有功功率和无功功率分布的概率模型。基于已知分布模型的密度估计和参数辨识的方法都属于基于模型的参数估计法,参数模型的优势在于其灵活性且具有数学表达式,但对于随机性、冲击性和复杂性较强的牵引负荷,参数估计法建立牵引负荷的概率模型时较难得到满意的结果[8-9]。而非参数估计方法对数据分布不依据任何假定,是一种完全基于数据驱动机制来研究数据分布特征的方法[10]。文献[11-13]采用核密度估计这种非参数估计方法建立了牵引负荷的概率模型。直方图是最早、最简单直观、应用最广泛的一种经典非参数密度估计方法[14-15]。由于直方图表征的是观测频率和相对频率的视觉信息,即密度函数的本质[16]。无论是参数估计还是非参数估计方法,都会将概率密度直方图作为其评价所建概率模型效果的参考基准,建立最佳的概率直方图至关重要。但针对牵引负荷的概率密度直方图的研究较少。

直方图构造受"箱宽"(或子空间宽度)和箱宽边界位置这2个因素的影响,箱宽边界位置的选择可得到较好的解决,最关键的是箱宽选择问题[1,15]。采用不同大小的箱宽(或不同数量的箱体),直方图概率密度形状会受到很大的影响。直方图最优平滑箱宽的选择是方差和偏差的折中,仍然是一个有难度的工作[15-16]。由于线路条件、车型、车速等很多因素的影响,不同牵引变电所的牵引负荷表现出不同的概率特征。直方图提供了对真实潜在概率密度的一致估计[17]。采用概率密度直方图建立牵引负荷的概率分布时,如何选择最优平滑箱宽,得到方差和偏差最佳折中的普适性方法是关键。求取直方图箱宽大小的常用方法可分为经验规则[17-19]和基于渐近估计的方法[20-21]。现有的这些方法的主要缺点是它们的渐近性质不能保证其较好的性能[17]。

本文提出一种基于最优箱宽建立牵引变电所负荷概率密度的直方图方法。利用Poisson点事件构造直方图,通过平均积分平方误差(MISE)定义价值函数,结合Sturges规则和平均移位直方图(ASH)的思想,计算最优箱宽。以3种不同采样间隔的牵引变电所负荷测试数据为例,通过概率密度直方图形状和后验检验的评价指标比较结果,验证了所提方法的有效性、适应性和稳健性。

收稿日期:2020-05-26;**修回日期**:2020-10-14
基金项目:甘肃省高等学校创新能力提升项目(2019B-049);甘肃省高等学校创新基金资助项目(2020A-036);兰州交通大学青年科学研究基金资助项目(20200013)
Project supported by the Innovative Ability Enhancement Project of Gansu Provincial Higher Education(2019B-049), Gansu University Innovation Fund Project(2020A-036) and the Science Foundation for Young Scholars of Lanzhou Jiaotong University(20200013)

1　直方图密度估计及箱宽选择问题

1.1　直方图的定义

直方图密度估计是通过样本数据构造概率密度的非参数估计经典方法。由于直方图反映了概率密度的本质，对于了解某一数据集分布的一般特征（如形状、变化趋势或位置）或建议可能的概率模型，直方图都是一种很好的方法。将随机变量取值范围的实轴区域划分为一些大小相等的子空间（或箱），每个子空间上估计的概率密度高度作为从端点到底边所形成的箱（或条形框），则得到一些由直立的箱排在一起而构成的直方图。概率密度直方图是归一化的直方图，为了估计真实的概率密度，希望具有一个非负且满足约束条件的密度函数 $\hat{f}(x)$。

$$\int \hat{f}(x)\mathrm{d}x = 1 \tag{1}$$

对于一维特征空间，已知 n 个随机样本 X_1、X_2、…、X_n，给出起点 x_0 和箱宽 h，将特征空间分割为不相交的箱体（或子区间）。令 $B_j=[t_j, t_{j+1}](j\in\mathbf{Z})$ 表示第 j 个箱，则 $t_{j+1}-t_j=h$，其中 $t_j=x_0+jh$。用 k_j 表示落入 B_j 中观测值的数量，则在点 x 处的一维概率密度直方图定义为：

$$\hat{f}_h(x)=\frac{k_j}{nh}=\frac{1}{nh}\sum_{i=1}^{n} I_{B_j}(X_i) \tag{2}$$

$$I_{B_j}(X_i)=\begin{cases}1 & X_i\in B_j\\ 0 & X_i\notin B_j\end{cases} \tag{3}$$

其中，$I_{B_j}(X_i)$ 为指标函数。

可见，要估计给定 x 的概率密度值，可以通过将样本数据中与 x 属于同一箱中的观测数 k_j 乘以 $1/(nh)$ 来得到 $\hat{f}_h(x)$ 的值。

由概率密度函数的定义可知，随机变量 x 落入 B_j 中的概率为：

$$P\{x\in B_j\}=P\{t_j<x\leqslant t_{j+1}\}=\int_{B_j}\hat{f}(x)\mathrm{d}x=\int_{t_j}^{t_{j+1}}\hat{f}(x)\mathrm{d}x \tag{4}$$

因此，$P\{x\in B_j\}$ 是概率密度函数 $\hat{f}(x)$ 平滑了（或取均值）的另一种形式，亦可通过估计 $P\{x\in B_j\}$ 来估计 $\hat{f}(x)$。

假设 X_1、X_2、…、X_n 独立同分布，则 n 个随机样本中有 k_j 个样本落入 B_j 中的概率服从二项分布[17]：

$$P_{k_j}=\mathrm{C}_n^{k_j}p^{k_j}(1-p)^{n-k_j}\quad k_j=0,1,\cdots,n;0<p<1 \tag{5}$$

1.2　Poisson 点事件构造直方图

当随机样本数量 n 很大时，二项概率的计算非常繁琐。由式（3）可知，X_i 落在 B_j 中的指标函数 $I_{B_j}(X_i)$ 的值为 1 或 0。因此可以使用 Poisson 定理近似计算二项分布的概率[22]：

$$\lim_{n\to+\infty}\mathrm{C}_n^{k_j}p_n^{k_j}(1-p_n)^{n-k_j}=\frac{\lambda^{k_j}}{k_j!}\mathrm{e}^{-\lambda} \tag{6}$$

其中，当 $n\to+\infty$ 时，$np_n\to\lambda$。

由式（6）可知，n 个随机样本中有 k_j 个样本落入 B_j 中的概率可近似服从 Poisson 分布。Poisson 分布的期望和方差均为 λ。

在经典密度估计中，样本容量是固定的。在 Poisson 分布假设下，样本容量可以不是固定的。因此，由 Poisson 点事件构造的直方图比由具有固定数量的样本构造的直方图具有更大的可变性。这使得在 Poisson 点过程假设下可以选择更宽泛的直方图最优箱宽大小。随着样本数据量的增加，固定和不固定样本容量之间的差异可以忽略不计。

1.3　直方图箱宽问题

由式（2）可知，构造直方图必须同时解决 2 个问题：箱宽的选择和箱宽边界位置的选择[1,23]。这 2 个选择中的每一个都会对所得的直方图产生较大的影响。关于箱宽边界位置选择的问题，可以通过箱体边界的移位对几个直方图求平均值来解决[24-25]。箱宽的选择主要是为了控制直方图应用于数据的“平滑”量，通常被称为平滑参数。较小的箱宽会导致不平滑的直方图，而较大的箱宽会使直方图过度平滑。

求取箱宽大小的常用经验规则有 Scott 规则、Freedman-Diaconis 规则（简称 FD 规则）、Sqrt 规则和 Sturges 规则等[15]。经验规则简单，但是不针对任何最优属性。经验规则是通过假设估计密度的特定特性而得出的，因此具有一定的局限性。例如 Scott 规则和 FD 规则更适用于单峰的正态分布，对多峰分布是次优的。更复杂的规则是基于对风险渐近估计的最小化，这是交叉验证情形，或者基于对真实密度函数平滑假设下的最优箱宽估计方法。

2　基于最优箱宽的直方图密度估计

2.1　最优箱宽计算方法

由式（2）可知，可用样本 X_i 在每个箱宽 h 中的计数频率来构造直方图，直方图最优平滑箱宽的选择是方差和偏差的折中。直方图概率密度估计 $\hat{f}_h(x)$ 能否较好地估计未知概率密度函数 $f(x)$，可以使用 MISE 进行评价[20]：

$$\delta_{\mathrm{MISE}}(\hat{f}_h)=E\left(\int(\hat{f}_h(x)-f(x))^2dx\right) \tag{7}$$

其中，E 为对观测数据不同概率密度的期望。虽然无法直接计算 MISE，但可通过样本数据估计 MISE。可将估计最优箱宽的大小转化为估计 MISE 的最小值，具有最小 MISE 的箱宽直方图是最佳的。

将 X_1、X_2、…、X_n 的区域 R 划分为宽度均是 h 的 N_h 个箱（或子区间），k_i 表示 x 在相同箱宽 h_i 中 X_i 的数量，则第 i 个箱宽的直方柱高度为 $k_i/(nh_i)$，其中 $h_i=h$。给定 h，$x\in[0,h]$ 的直方柱高度预期值 θ 为概率密度函数 $f(x)$ 的平均值，即：

$$\theta=\frac{1}{h}\int_0^h f(x)\mathrm{d}x \tag{8}$$

θ的无偏估计为$\hat{\theta}=k/(nh)$，这是$x\in[0,h]$直方柱的经验高度，其中k为落入$[0,h]$中观测值的数量。

由$f(x)$在$x\in[0,h]$中的平均值，式(7)可表示为：

$$\delta_{\mathrm{MISE}}(\hat{f}_h)=\frac{1}{h}\int_0^h E\left((\hat{\theta}-f(x))^2\right)\mathrm{d}x \tag{9}$$

$\delta_{\mathrm{MISE}}(\hat{f}_h)$进一步可以分解为两部分：

$$\delta_{\mathrm{MISE}}(\hat{f}_h)=E\left((\hat{\theta}-\theta)^2\right)+\frac{1}{h}\int_0^h(f(x)-\theta)^2\,\mathrm{d}x \tag{10}$$

式(10)第一项为样本误差，表示概率密度估计值$\hat{\theta}$的随机波动；第二项为系统误差，表示在$x\in[0,h]$内概率密度函数$f(x)$在其平均值附近的平均波动。

用$\bar{\theta}$表示$f(x)$的平均值，如图1所示。由于未知$f(x)$与h独立，则式(10)进一步分解为：

$$\delta_{\mathrm{MISE}}(\hat{f}_h)=E\left((\hat{\theta}-\theta)^2\right)+\frac{1}{R}\int_0^R(f(x)-\bar{\theta})^2\,\mathrm{d}x-(\theta-\bar{\theta})^2 \tag{11}$$

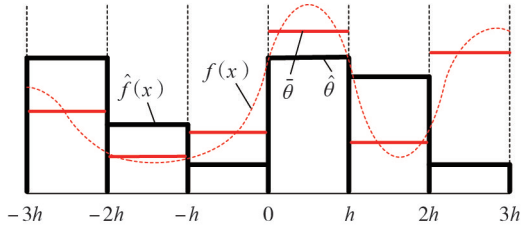


图1 直方图各符号含义的示意图

Fig.1 Schematic diagram of meaning of each symbol in histogram

由$\delta_{\mathrm{MISE}}(\hat{f}_h)$减去式(11)中的第二项，可定义价值函数如下：

$$C_n(h)=\delta_{\mathrm{MISE}}(\hat{f}_h)-\frac{1}{R}\int_0^R(f(x)-\bar{\theta})^2\,\mathrm{d}x=$$
$$E\left((\hat{\theta}-\theta)^2\right)-(\theta-\bar{\theta})^2 \tag{12}$$

该价值函数$C_n(h)$是通过直接分解MISE得到的风险函数，其第一项为样本误差；第二项为理想直方图的方差，它是未知的。预期均值θ不能直接得到，但可用观测估计值$\hat{\theta}$替换θ。采用无偏估计量$E(\hat{\theta})=\theta$的分解规则，则有：

$$E\left((\hat{\theta}-E(\hat{\theta}))^2\right)=E\left((\hat{\theta}-\theta)^2\right)+(\theta-\bar{\theta})^2 \tag{13}$$

其中，$E\left((\hat{\theta}-E(\hat{\theta}))^2\right)$为所求直方图的方差。

从而，可将价值函数$C_n(h)$化简为：

$$C_n(h)=2E\left((\hat{\theta}-\theta)^2\right)-E\left((\hat{\theta}-E(\hat{\theta}))^2\right) \tag{14}$$

由$\hat{\theta}=k/(nh)$，结合均值-方差关系得：

$$E\left((\hat{\theta}-\theta)^2\right)=E(\hat{\theta})/(nh) \tag{15}$$

由式(6)可知，n个随机样本中有k个样本落入箱宽h中的概率可近似服从Poisson分布，则有$P_k=(\theta nh)^k\mathrm{e}^{-\theta nh}/k!$，其期望和方差均为$\theta nh$。设样本数总量为$N_{\mathrm{t}}$，从而有：

$$C_n(h)=\frac{2}{nh}E(\hat{\theta})-E\left((\hat{\theta}-E(\hat{\theta}))^2\right)=\frac{2\bar{K}-V}{(nh)^2} \tag{16}$$

$$\bar{K}=\frac{1}{N_{\mathrm{t}}}\sum_{i=1}^{N_{\mathrm{t}}}k_i \tag{17}$$

$$V=\frac{1}{N_{\mathrm{t}}}\sum_{i=1}^{N_{\mathrm{t}}}(k_i-\bar{k})^2 \tag{18}$$

由式(16)可知，由2倍样本误差减去所求直方图方差所表示的价值函数$C_n(h)$，转化成了价值函数$C_n(h)$由样本均值$\bar{K}$、样本方差V、样本量N_{t}和箱宽h所决定的关系。求相邻价值函数差值$\Delta C_n(h)=C_{n_v}(h)-C_{n_v+1}(h)$的最小值，其中$v=1,2,\cdots,N_{\mathrm{t}}-1$，可得到最优箱宽$h^*$：

$$h^*=\arg\min_h \Delta C_n(h) \tag{19}$$

可见，在相邻价值函数差值最小的情况下，本文所提方法得到的最优箱宽由样本均值、样本方差和样本量决定。从而最优箱数N_h^*为：

$$N_h^*=\left[R/h^*\right] \tag{20}$$

其中，[ ]表示取整，可通过N_h^*分析箱数对概率密度直方图的影响。

2.2 选择最优箱宽的步骤

为了降低计算复杂度，本文引入Sturges规则[16,26]，通过初始箱数定义箱宽矢量，可以提高寻优速度。而且，利用ASH的思想[24-25]建立移位等分箱宽，不仅可以避免直方图原点选择问题，而且可以在分层的箱宽矢量中找到更精细的最优箱宽。

求得最优箱宽后，可以通过式(2)得到概率密度直方图。求解最优箱宽h^*的具体步骤如下。

(1)建立箱宽矢量。剔除异常数据，对样本观测值X_1、X_2、…、X_n的非重复值进行排序。设$R=X_{\max}-X_{\min}$。采用最常用的Sturges规则，计算初始箱数$N_0=1+\log_2 n$。令箱数向量$\boldsymbol{N}=[2,3,\cdots,N_0]$，则箱宽向量$\boldsymbol{h}=R/\boldsymbol{N}=[R/2,R/3,\cdots,R/N_0]$。

(2)计算移位等分箱宽。对箱宽向量$\boldsymbol{h}$空间中每个箱宽h_{V}构造支集$S_i=\{[0,h_i)\}$，其中$i=1,2,\cdots,N_0-1$。令$\delta_i=h_i/M\,(M>0)$，对每个支集S_i进行移位等分为M个子区间以得到更精细的网格$B_{i,j}=[j\delta_i,(j+1)\delta_i)$，其中$i=1,2,\cdots,N_0-1$且$j=0,1,\cdots,M-1$。如，将$[0,h_i)$分为$B_{i,0}=[0,\delta_i)$、$B_{i,1}=[\delta_i,2\delta_i)$、…、$B_{i,M-1}=[(M-1)\delta_i,M\delta_i)=[(M-1)\delta_i,h_i)$，那么有$M$个覆盖$B_{i,0}$的移位等分箱宽$h_{ij}=j\delta_i$。取移位等分箱宽的区间数$M=200$。

(3)对x落入箱宽子区间$B_{i,j}$的样本观测值进行计数，即频数k_{ij}。

(4)根据式(17)和式(18)分别计算每个箱宽中频数k_{ij}的期望和方差。

(5)根据式(16)计算价值函数。

（6）求解最优箱宽，并重复步骤（3）—（5）。由式（19）得到最优箱宽 h^*。

2.3　后验检验

计算箱宽大小常用的经验规则方法计算公式见表1，利用这些经验规则方法验证本文所提的基于最优箱宽直方图建立牵引变电所负荷概率密度的优劣性。表中，$\hat{\sigma}$ 为样本方差；I_Q 为样本的3/4分位值与1/4分位值的差额。为了对本文所提的最优箱宽计算方法和已有计算箱宽的规则方法进行比较，并评价直方图表征牵引负荷概率分布的性能，采用后验检验定量评估直方图与数据观测分布之间的差异。

表1　求取箱宽常用规则方法的计算公式

Table 1　Calculation formulas of bin width for commonly used rules

规则方法	计算公式	规则方法	计算公式
Scott规则	$3.5\hat{\sigma}N_t^{-1/3}$	Sqrt规则	$\sqrt{N_t}$
FD规则	$2I_Q N_t^{-1/3}$	Sturges规则	$1+\log_2 N_t$

直方图构造的是离散的密度估计，而核密度估计则等效于将直方图采用核函数进行平滑。因此将采用高斯核函数的核密度估计所得到的牵引负荷概率密度作为基准值。然后，利用三次样条插值获取各种箱宽计算方法所得直方图与核密度估计基准值相对应的概率值。采用均方根误差（RMSE）和平均误差百分比（MAPE）进行牵引负荷直方图概率密度的后验检验，分别如式（21）和式（22）所示。RMSE和MAPE越小，说明牵引负荷的直方图概率密度与数据观测分布之间的差异越小。

$$\delta_{\mathrm{RMSE}}=\sqrt{\frac{1}{N_h}\sum_{i=1}^{N_h}(\hat{R}_{\mathrm{h}i}-R_{\mathrm{k}i})^2} \tag{21}$$

$$\delta_{\mathrm{MAPE}}=\frac{1}{N_h}\sum_{i=1}^{N_h}\left|\frac{\hat{R}_{\mathrm{h}i}-R_{\mathrm{k}i}}{R_{\mathrm{k}i}}\right|\times 100\% \tag{22}$$

其中，$\hat{R}_{\mathrm{h}i}$、$R_{\mathrm{k}i}$ 分别为牵引负荷的直方图和核密度估计概率分布在第 i 个子区间的概率密度值。

为全面评估各种箱宽计算方法下直方图概率密度的精度，定义综合评估指标（CTI）为：

$$\delta_{\mathrm{CTI}}=\frac{\delta_{\mathrm{RMSE}}}{\delta_{\mathrm{RMSE,max}}}+\frac{\delta_{\mathrm{MAPE}}}{\delta_{\mathrm{MAPE,max}}} \tag{23}$$

其中，$\delta_{\mathrm{RMSE,max}}$ 和 $\delta_{\mathrm{MAPE,max}}$ 分别为RMSE和MAPE的最大值。CTI越小，说明直方图表征牵引负荷概率分布的效果越好。

3　算例分析

3.1　牵引变电所测试数据

通过3种不同采样间隔的牵引负荷测试数据，验证本文所提的基于最优箱宽直方图密度估计方法的有效性、正确性和适应性。在固定的测试周期中，不同的采样间隔可以得到不同的测试数据样本量，所反映的牵引负荷特征也会不尽相同。这3个牵引负荷的测试数据参数如表2所示。应当说明，空载情况下的牵引负荷分布是一个尖峰，对建立的牵引负荷概率模型非常不利，因此需先将空载牵引负荷忽略，只考虑带电情形[3]。由空载时的样本量除以牵引负荷在一个周期的总样本量就可确定空载率。

表2　牵引负荷测试数据参数

Table 2　Parameters of traction load test data

序号	牵引负荷数据	样本量	样本极差
1	Case 1（采样间隔为250 ms）	186498（去除空载数据后）	557.50
2	Case 2（采样间隔为3 s）	9007（仅牵引工况部分）	21.86
3	Case 3（采样间隔为10 s）	8640	18.52

Case 1：某普速电气化铁路运行的交直型和交直交型电力机车的功率因数分别为0.81、0.97以上，额定电流分别为175、350 A。采用Fluke 1760电能质量测试仪对某一牵引变压器的馈线电流进行测试，测试周期为24 h，采样间隔为250 ms。馈线电流时序曲线如图2（a）所示，可看出馈线电流具有明显的波动性、间歇性和较高的空载率。将去除空载数据后的馈线电流测试数据记为Case 1。

Case 2：对某高速铁路的一个牵引变电所采用Fluke 1760电能质量测试仪在220 kV三相侧变压器对有功功率进行测试，测试周期为24 h，采样间隔为3 s。有功功率的时序曲线如附录中图A1（a）所示，可看出其除了具有较强的波动性、间歇性和较高的空载率，还具有一定的规律性。将去除空载数据后的该有功功率测试数据记为Case 2。

Case 3：某220 kV变电所在给某高速铁路牵引站供电的同时，也给电力变压器提供电源。牵引变压器容量为25 MV·A，电气参数为（220±2×2.5%）kV／（2×27.5）kV。电力变压器容量为4 MV·A，电气参数为（220±2.5%）kV／10 kV。采用Fluke 1760电能质量测试仪在220 kV三相侧变压器对有功功率进行测试，测试周期为24 h，采样间隔为10 s。有功功率的时序曲线见附录中图A2（a），由于含有电力变压器的负荷的间歇性和空载率不是很明显，因此不去除空载数据。将该有功功率测试数据记为Case 3。

3.2　最优箱宽和最优箱数结果

采用本文的最优箱宽直方图密度估计和表1所示的4种经验规则方法，计算Case 1、Case 2和Case 3下牵引负荷概率密度直方图的最优箱宽和最优箱数如表3所示。对于Scott、FD和Sqrt规则所得的最优箱宽和最优箱数，Case 1的结果明显大于Case 2和Case 3的，而Case 2和Case 3的结果很相近。对于Sturges规则所得的最优箱数，3个算例下的结果都很接近，而且箱数都在20以内。这是由于样本极差

不同，通过最优箱宽h^*得到的最优箱数N_h^*有较大的区别。箱数过多，直方图概率密度就会过于集中，这会导致直方图概率密度不平滑性；箱数过少，直方图概率密度就会过于平滑，而掩盖样本信息。结合式(16)和式(19)，本文方法所得的最优箱宽是在相邻价值函数差值最小的情况下，由样本均值、样本方差和样本量所决定。由表2可知，不同牵引负荷测试数据的样本量和样本极差有较大的不同。由表3可知，本文方法所得的最优箱宽和最优箱数对不同的采样间隔、样本量和样本极差具有一定的稳健性。

表3 直方图的最优箱宽和最优箱数

Table 3 Optimal bin width of histogram and optimal number of bin

算例	指标	数值				
		Scott规则	FD规则	Sqrt规则	Sturges规则	本文方法
Case 1	h^*	3.6611	1.9516	1.2905	29.3421	12.8336
	N_h^*	152	285	432	19	44
Case 2	h^*	0.5343	0.4642	0.2301	1.4574	1.4283
	N_h^*	40	47	95	15	16
Case 3	h^*	0.4070	0.3235	0.1991	1.2343	0.6437
	N_h^*	45	57	93	15	29

3.3 牵引负荷最优直方图

从图2(a)、图A1(a)和图A2(a)可以看出，这3个算例的牵引负荷时序曲线各具特点。以核密度估计所得的牵引变电所负荷概率密度作为基准，采用本文所提方法和表1所示的4种经验规则方法，分别建立Case 1、Case 2和Case 3牵引负荷的概率模型。

(1)Case 1的牵引负荷最优直方图。

图2(b)为Scott、FD、Sqrt、Sturges规则和本文方法所得的Case 1馈线电流的直方图概率密度。由图可见，FD和Sqrt规则直方图具有分布过度集中的现象，所表征的概率密度体现了过多的细节而不够平滑；Sturges规则直方图只保留了1个波峰，相比其他方法直方图具有一高一低2个波峰，其所表征的概率密度过于平坦，过度丢失了Case 1馈线电流的概率分布信息；Scott规则和本文方法直方图能较好地直观看出Case 1馈线电流概率分布情况，但Scott规则直方图更加密集而平滑性较差。因此，本文方法所建立的Case 1馈线电流直方图在概率密度形状方面优于表1中的4种规则方法。

(2)Case 2的牵引负荷最优直方图。

附录中图A1(b)为Scott、FD、Sqrt、Sturges规则和本文方法所得的Case 2有功功率的直方图概率密度。由图可见，除了Sqrt规则直方图存在明显的过度集中而不平滑外，Scott、FD、Sturges规则和本文方法所得的直方图都能直观地表征出Case 2牵引工况下有功功率的概率密度，但仍有一定的区别。从表3可以看出，Scott和FD规则得到的最优箱宽和最优箱数比较接近，而Sturges规则和本文方法得到的最优箱宽和最优箱数很接近。图A1(b)中，Scott和FD规则所得的直方图概率密度较相似，而Sturges规则和本文方法所得的直方图概率密度较相似。

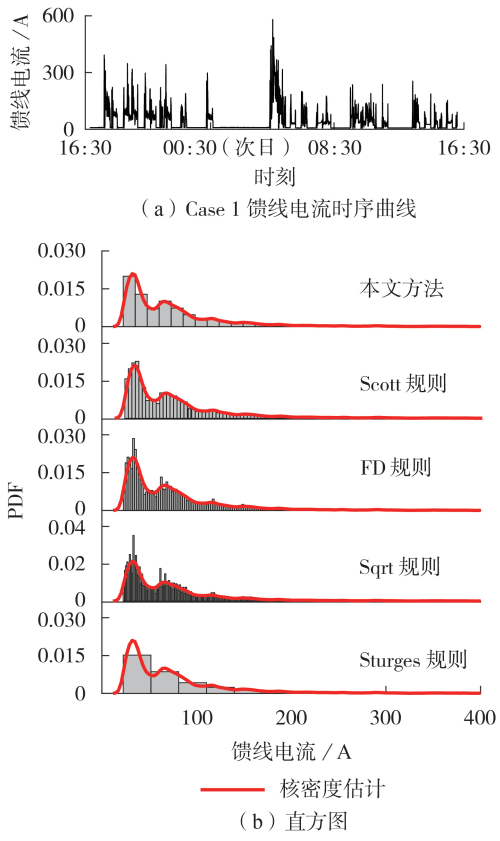


图2 Case 1牵引负荷及其最优箱宽直方图

Fig.2 Traction load of Case 1 and its optimal bin width histogram

(3)Case 3的牵引负荷最优直方图。

附录中图A2(b)为Scott、FD、Sqrt、Sturges规则和本文方法所得的Case 3有功功率的直方图概率密度。由于Case 3未去除空载牵引负荷，各直方图在空载值处存在明显的尖峰。从直方图概率密度的形状来看，Sqrt、Sturges规则所得的有功功率直方图概率密度分别表现为不平滑和过度光滑；Scott规则和本文方法直方图比较接近；FD规则直方图能够描述Case 3的概率分布，但在0值处的概率密度明显高于Scott规则和本文方法直方图。

综上可见，Sqrt规则总是存在概率密度过度集中的缺点；Sturges规则可能会存在概率密度过于平滑而丢失样本数据信息的缺点；Scott规则和FD规则得到的直方图概率分布可以接受，但可能存在密集现象；本文方法得到的直方图均能较好地表征出这3种牵引负荷不同特性的概率分布，可见本文方法描述不同的牵引负荷概率分布具有较好的适应性。

3.4 后验检验结果

由2.3节中后验检验的比较方法，对于子区间数分别为200和400的后验检验指标结果如表4所示。

表4　后验检验结果

Table 4　Results of posterior tests

M	计算方法	Case 1			Case 2			Case 3		
		RMSE	MAPE	CTI	RMSE	MAPE	CTI	RMSE	MAPE	CTI
200	Scott规则	0.0047	7.46×10^{3}	0.0616	0.8693	4.97×10^{4}	0.2581	0.0161	1.01×10^{5}	0.6002
	FD规则	0.0484	7.96×10^{4}	0.6479	0.9957	5.22×10^{4}	0.2874	0.0159	4.15×10^{4}	0.4265
	Sqrt规则	0.1217	3.18×10^{5}	2.0000	5.0759	5.72×10^{5}	2.0000	0.0513	3.55×10^{5}	2.0000
	Sturges规则	0.0030	1.08×10^{3}	0.0278	0.0342	3.88×10^{3}	0.0135	0.0165	8.23×10^{4}	0.5545
	本文方法	0.0018	1.15×10^{3}	0.0184	0.0215	3.29×10^{3}	0.0100	0.0111	4.76×10^{4}	0.3496
400	Scott规则	0.0040	4.61×10^{3}	0.0624	0.8332	4.08×10^{4}	0.2602	0.0157	8.34×10^{4}	0.6092
	FD规则	0.0422	4.83×10^{4}	0.6543	0.9537	4.29×10^{4}	0.2897	0.0158	3.41×10^{4}	0.4404
	Sqrt规则	0.1047	1.92×10^{5}	2.0000	4.8227	4.67×10^{5}	2.0000	0.0490	2.89×10^{5}	2.0000
	Sturges规则	0.0029	7.20×10^{2}	0.0312	0.0328	3.18×10^{3}	0.0136	0.0161	6.90×10^{4}	0.5681
	本文方法	0.0018	7.50×10^{2}	0.0211	0.0212	2.70×10^{3}	0.0102	0.0109	3.94×10^{4}	0.3596

Sqrt规则的CTI值均为2，由于RMSE和MAPE的最大值均受Sqrt规则的RMSE和MAPE的影响，而Scott、FD、Sturges规则和本文方法的CTI值随着区间数的增加而有所变大。除了Case 1中Sturges规则的MAPE最小（略小于本文方法的MAPE），对于Case 2和Case 3，本文方法的MAPE均是最小的。

3个算例下，相比Scott、FD、Sqrt和Sturges规则，本文方法得到的RMSE和CTI值均是最小的。这说明本文方法的最优直方图所描述的牵引变电所负荷概率分布与数据观测分布（即核密度估计分布）之间的差异最小。对照区间数M分别为200和400的后验检验指标，可看出4种规则和本文方法的RMSE和MAPE均随着区间数的增大而降低，这说明考虑的样本信息越多，得到的精度越高。该结果验证了本文所提最优箱宽直方图密度估计方法的有效性和准确性。图2、图A1和图A2的比较结果也表明了本文方法所得的最优直方图能较好地表征不同类型牵引负荷的概率特性，具有较好的适应性。

4　结论

采用传统直方图表征的牵引负荷概率密度极易受箱宽（或箱数）大小的影响，本文提出了基于最优箱宽直方图建立牵引负荷概率密度的方法。利用Poisson点事件构造直方图，通过MISE定义价值函数，结合Sturges规则和移位等分箱宽方法，在相邻价值函数差值最小的情况下，本文方法所得的最优箱宽由样本均值、样本方差和样本量决定。Sturges规则定义初始箱数和箱宽矢量，可降低复杂度和提高寻优速度。移位等分箱宽方法不仅可以避免直方图原点选择问题，而且可以在分层的箱宽矢量中找到更精细的最优箱宽。通过3种不同采样间隔的牵引变电所负荷测试数据，以Scott、FD、Sqrt和Sturges这4种经验规则作为比较方法，以核密度估计所得的牵引负荷概率密度分布作为基准，通过RMSE、MAPE及CTI等指标后验检验，结果验证了基于最优箱宽直方图密度估计方法的有效性、适应性和稳健性。所提方法也可扩展至其他领域的直方图概率建模。

附录见本刊网络版（http://www.epae.cn）。

参考文献：

[1] SILVERMAN B W. Density estimation for statistics and data analysis[M]. London, UK: Chapman and Hall, 1986: 1-33.

[2] SUZUKI S, BABA J, SHUTOH K, et al. Effective application of Superconducting Magnetic Energy Storage (SMES) to load leveling for high speed transportation system[J]. IEEE Transactions on Appiled Superconductivity, 2004, 14(2): 713-716.

[3] 张丽艳，李群湛，解绍锋，等. 牵引变电所馈线电流的概率模型[J]. 西南交通大学学报，2009，44(6)：848-854.
ZHANG Liyan, LI Qunzhan, XIE Shaofeng, et al. Probability distribution of feeder current of traction substation[J]. Journal of Southwest Jiaotong University, 2009, 44(6): 848-854.

[4] 唐开林，李群湛，张丽艳，等. 电气化铁道牵引网馈线电流概率分布[J]. 电力系统及其自动化学报，2010，22(6)：12-16.
TANG Kailin, LI Qunzhan, ZHANG Liyan, et al. Probability distribution of feeder current of electrified railway traction [J]. Proceedings of the CSU-EPSA, 2010, 22(6): 12-16.

[5] 解绍锋，李群湛，赵丽平. 电气化铁道牵引负载谐波分布特征与概率模型研究[J]. 中国电机工程学报，2005，25(16)：79-83.
XIE Shaofeng, LI Qunzhan, ZHAO Liping. Study on harmonic distribution characteristic and probability[J]. Proceedings of the CSEE, 2005, 25(16): 79-83.

[6] 杨少兵，吴命利. 基于实测数据的高速动车组谐波分布特性与概率模型研究[J]. 铁道学报，2010，32(3)：33-38.
YANG Shaobing, WU Mingli. Study on harmonic distribution characteristics and probability model of high speed EMU based on measured data[J]. Journal of the China Railway Society, 2010, 32(3): 33-38.

[7] 杨少兵，吴命利. 电气化铁道牵引变电所负荷概率模型[J]. 电力系统自动化，2010，34(24)：40-45.
YANG Shaobing, WU Mingli. A load probability model for electrified railway traction substations[J]. Automation of Electric Power Systems, 2010, 34(24): 40-45.

[8] 刘晓楠，周介圭，贾宏杰，等. 基于非参数核密度估计与数值天气预报的风速预测修正方法[J]. 电力自动化设备，2017，37(10)：15-20.
LIU Xiaonan, ZHOU Jiegui, JIA Hongjie, et al. Correction method of wind speed prediction based on non-parametric kernel density estimation and numerical weather prediction[J]. Electric Power Automation Equipment, 2017, 37(10): 15-20.

[9] 李亚琼，周胜军，王同勋，等. 基于优选带宽核密度估计的谐波概率潮流分析方法[J]. 电力自动化设备，2017，37(8)：131-136.

LI Yaqiong, ZHOU Shengjun, WANG Tongxun, et al. Harmonic probabilistic power-flow analysis based on kernel density estimation with optimized bandwidth design[J]. Electric Power Automation Equipment, 2017, 37(8): 131-136.
[10] TSYBAKOV A B. Introduction to nonparametric estimation [M]. New York, USA: Springer-Verlag, 2009: 2-19.
[11] 陈民武，刘洋，韩旭东，等. 高速铁路牵引负荷谐波分布的非参数估计模型与预测评估[J]. 电网技术，2017，41(8)：2598-2603.
CHEN Minwu, LIU Yang, HAN Xudong, et al. Harmonic distribution modeling based on non-parametric estimation and harmonic prediction evaluation for high speed railway traction load[J]. Power System Technology, 2017, 41(8): 2598-2603.
[12] 张丽艳，陈映月，韩正庆. 基于改进聚类方式的牵引负荷分类方法[J]. 西南交通大学学报，2020，55(1)：27-33，40.
ZHANG Liyan, CHEN Yingyue, HAN Zhengqing. Traction load classification method based on improved clustering method[J]. Journal of Southwest Jiaotong University, 2020, 55(1): 27-33, 40.
[13] CHE Yulong, WANG Xiaoru, LÜ Xiaoqin, et al. Study on probability distribution of electrified railway traction loads based on kernel density estimator via diffusion[J]. International Journal of Electrical Power and Energy Systems, 2019, 106: 383-391.
[14] 牛君. 基于非参数密度估计点样本分析建模的应用研究[D]. 济南：山东大学，2007.
NIU Jun. Application research based on non-parametric density estimation point sample analysis modeling[D]. Jinan: Shandong University, 2007.
[15] KNUTH K H. Optimal data-based binning for histograms and histogram-based probability density models[J]. Digital Signal Processing, 2019, 95: 1-16.
[16] SCOTT D W. Averaged shifted histogram[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2(2): 160-164.
[17] BIRGE L, ROZENHOLC Y. How many bins should be put in a regular histogram[J]. ESAIM Probability and Statistics, 2006, 10: 24-45.
[18] SCOTT D W. On optimal and data-based histograms[J]. Biometrika, 1979, 66(3): 605-610.
[19] RUDEMO M. Empirical choice of histograms and kernel density estimators[J]. Scandinavian Journal of Statistics, 1982, 9: 65-78.
[20] SHIMAZAKI H, SHINOMOTO S. A method for selecting the bin size of a time histogram[J]. Neural Computation, 2007, 19(6): 1503-1527.
[21] SCOTT D W. Averaged shifted histograms: effective nonparametric density estimators in several dimensions[J]. Annals of Statistics, 1985, 13(3): 1024-1040.
[22] 王丽霞. 概率论与随机过程：理论、历史及应用[M]. 北京：清华大学出版社，2010：51-53.
[23] WAND M P, JONES M C. Kernel smoothing[M]. London, UK: Chapman and Hall, 1995: 5-7.
[24] BOUREL M, FRAIMAN R, GHATTAS B. Random average shifted histograms[J]. Computational Statistics & Data Analysis, 2014, 79: 149-164.
[25] FERNANDO T M K G, MAIER H R, DANDY G C. Selection of input variables for data driven models: an average shifted histogram partial mutual information estimator approach[J]. Journal of Hydrology (Amsterdam), 2009, 367(3-4): 165-176.

作者简介：

车玉龙

车玉龙(1988—)，男，甘肃武山人，讲师，博士研究生，主要研究方向为牵引负荷的不确定性建模及应用(**E-mail**：cylg717@163.com)；

王晓茹(1962—)，女，四川成都人，教授，博士研究生导师，主要研究方向为电力系统稳定性分析与控制(**E-mail**：xrwang@home.swjtu.edu.cn)；

吕晓琴(1980—)，女，四川成都人，副教授，博士，主要研究方向为电气化铁路车网耦合建模及分析(**E-mail**：xiaoqin93@163.com)。

(编辑 李莉)

Load probability modeling method of traction substation based on optimal bin width histogram

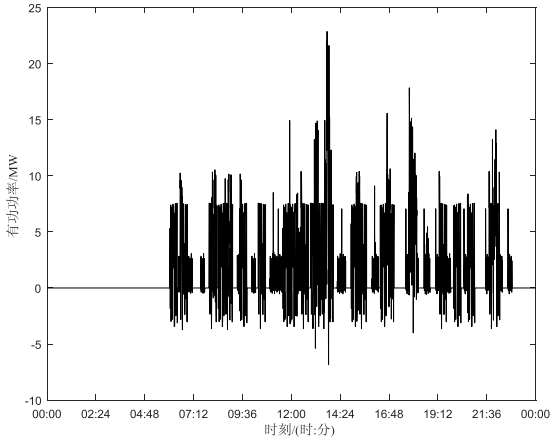
CHE Yulong[1,2], WANG Xiaoru[2], LÜ Xiaoqin[2], KANG Yongqiang[3]

(1. School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China;
2. School of Electrical Engineering, Southwest Jiaotong University, Chengdu 611756, China;
3. School of New Energy and Power Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China)

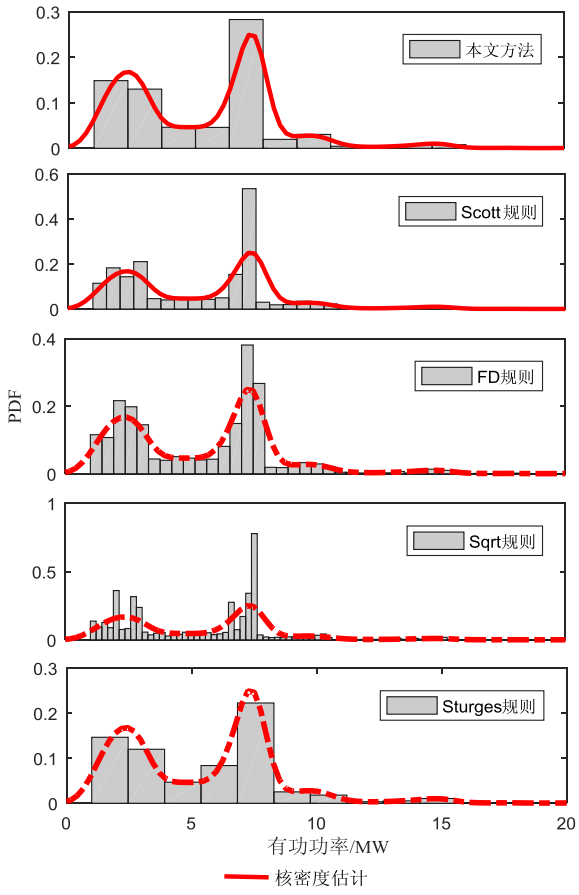
Abstract: As the traction load in electric railway is affected by many random factors, it is difficult to uniformly describe the load probability model by the known parameter estimation models. Histogram is the simplest, most intuitive and widely used non-parametric estimation method, but the shape of the probability density obtained from histogram is extremely susceptible to bin width (or bin number). In order to obtain the optimal histogram of traction load probability density, a histogram method based on the optimal bin width is proposed to build its probability model. The histogram is constructed by Poisson point events, the value function is defined by mean integrated squared error, and the optimal bin width is calculated by Sturges rule and the average shift histogram idea. Based on the load test data of traction substations with three different sampling intervals, through the shape of probability distribution and the posterior test, the validity, adaptability and robustness of the proposed method are verified by comparing with four traditional empirical rule methods for bin width calculation.

Key words: traction load; probability model; histogram; value function; optimal bin width; average shift

附录



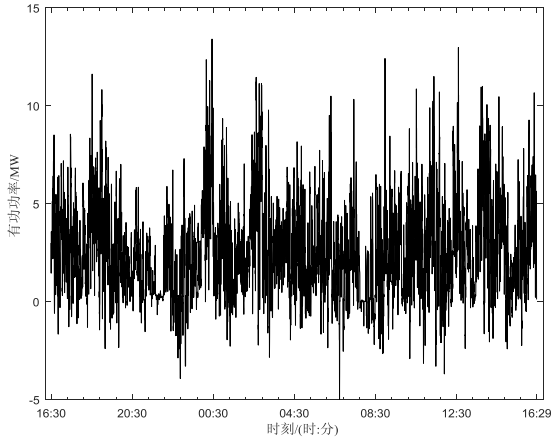
(a) Case 2 有功功率时序曲线



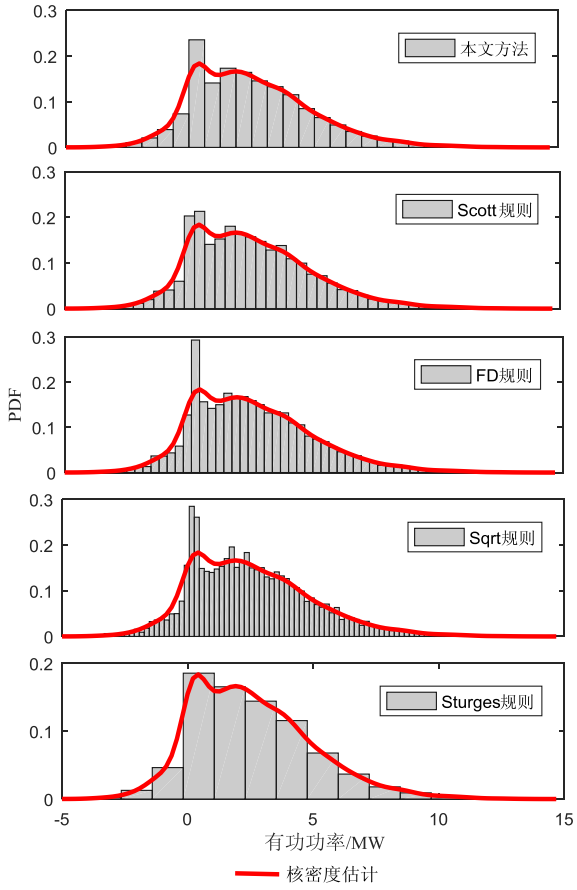
(b) 直方图

图 A1　Case 2 牵引负荷及其最优箱宽直方图

Fig.A1　Traction load of Case 2 and its optimal bin width histogram



(a) Case 3 有功功率时序曲线



(b) 直方图

图 A2　Case 3 牵引负荷及其最优箱宽直方图

Fig.A2　Traction load of Case 3 and its optimal bin width histogram