Electric Power Systems Research 121 (2015) 28–37

Contents lists available at ScienceDirect

Electric Power Systems Research

journal homepage: www.elsevier.com/locate/epsr

Estimating wind speed probability distribution by diffusion-based kernel density method

Xiaoyuan Xu, Zheng Yan*, Shaolun Xu

Key Laboratory of Control of Power Transmission and Conversion, Ministry of Education, Department of Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

ARTICLE INFO

Article history:
Received 16 July 2014
Received in revised form 15 November 2014
Accepted 28 November 2014

Keywords:
Diffusion partial differential equation
Discrete cosine transformation
Kernel density estimation
Parametric distribution model
Wind speed probability distribution
Goodness of fit test

ABSTRACT

Accurate estimation of wind speed probability distributions is a challenging task in wind power planning and operation. Different from the commonly used parametric methods which consist of selecting a suitable parametric model and estimating the parameters, this paper presents an improved non-parametric method to estimate wind speed probability distributions. Based on the diffusion partial differential equation in finite domain, this method accounts for both bandwidth selection and boundary correction of kernel density estimation. Preprocessing techniques are designed to handle data with different recording manners to produce smooth probability density functions. Probability densities of specific grid points are obtained by inverse discrete cosine transformation and are further used to calculate assessment indices of wind resources. The method has been tested to estimate probability densities of parametric distributions and actual wind speed data measured in different places. Simulation results show that the proposed approach is of practical value in fitting wind speed distribution models.

1. Introduction

As more and more wind power has been utilized around the world, it is of great value to estimate probability distributions of wind speed to maximize the efficiency of wind power generations. In power system planning and reliability analysis, most of the researches on wind speed probability distributions are based on parametric distribution models [1–6]. And the methods estimating wind speed probability distributions are usually parametric methods which consist of selecting a suitable parametric distribution and estimating the parameters [7–13]. The commonly used parametric distribution models can be divided into two categories:

(1) Unimodal parametric distributions including the Weibull distribution, the Gamma distribution, the Rayleigh distribution and so on.
(2) Multimodal parametric distributions, especially bimodal models, for example, the mixture Weibull distribution of two components, the mixture distribution of truncated Normal and Weibull.

The parametric methods are widely used for the efficiency in estimation and can be applied to estimate wind characteristics at the sites for which no wind data is available [14,15]. However, there are also challenges for the parametric methods. First, there is no rule in selecting the theoretical distribution and a distribution model which can represent wind regimes at some wind farms may not work well for others. Second, an estimated parametric model may not always be satisfactory results because of the extreme randomness of wind speed in both time and space. Thus, the non-parametric methods are introduced to estimate wind speed probability distributions [16,17]. Kernel density estimation (KDE) is one of the most popular non-parametric distribution estimating methods [18–20]. Up to now, most of the papers concentrate on the mathematical theories of KDE and only a few of them actually model the probability distributions of wind speed. A practical kernel density method is proposed in [16] to estimate long-time wind speed probability distributions. A smooth multivariate wind distribution model is developed in [17] to capture the coupled variation of wind speed, wind direction, and air density.

In this paper an improved diffusion-based kernel density method (DKDM) is presented to estimate wind speed probability distributions. DKDM accounts for both bandwidth selection and boundary correction of KDE and discrete cosine transformation is adopted in DKDM to reduce the computational complexity. In order to produce smooth probability densities, a uniformly distributed

* Corresponding author. Tel.: +86 13621931964.
E-mail addresses: xuxiaoyuansjtu@gmail.com (X. Xu), yanz@sjtu.edu.cn (Z. Yan), slxu@sjtu.edu.cn (S. Xu).

http://dx.doi.org/10.1016/j.epsr.2014.11.029
0378-7796/

random number is added to the original wind speed data and its value range depends on the corresponding recording frequency and resolution. The probability densities of specific grid points are obtained by inverse discrete cosine transformation and are further used to calculate assessment indices of wind resources. The proposed method is compared with other approaches in estimating probability densities of parametric distribution functions and actual wind speed data. The simulation results demonstrate the practicality of this method.

2. Diffusion-based kernel density method

2.1. Introduction of kernel density estimation

Suppose we have observed data $X_1, X_2, \ldots, X_N$ from a common distribution with the probability density function $f(x)$, and use the kernel density method to estimate the density as follows [18]:

$$\hat{f}(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - X_i}{h}\right) \tag{1}$$

where $K(\cdot)$ is the kernel function, h is the bandwidth and N is the size of data.

There are many kernel functions and the Gaussian kernel is selected as the kernel function in this paper:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) \tag{2}$$

It turns out that the choice of bandwidth is much more important than the choice of kernel functions for the behavior of estimating results. Small values of bandwidth make the estimation look "wiggly" and show spurious features, whereas big values of bandwidth will make the result over-smooth in a sense that it may not reveal the structural features, such as multimodality [19]. In addition, the original KDE assumes the domain of the density to be infinite. Thus, KDE suffers from boundary problems if the domain has finite endpoints [21,22].

There are two classes of methods to estimate the bandwidth of KDE: the cross-validation methods which try to look at $ISE(h)$, and the plug-in methods which try to minimize $MISE(h)$ [23].

$$ISE(h) = \int (\hat{f}_h(x) - f(x))^2 dx \tag{3}$$

$$MISE(h) = E[\int \{\hat{f}_h(x) - f(x)\}^2 dx] \tag{4}$$

where $\hat{f}_h(x)$ represents the probability density estimated by KDE, $f(x)$ represents the true density distribution function and $E(\cdot)$ is the expectation of variables.

(1) Cross-validation methods

$$ISE(h) = \int \{\hat{f}_h(x) - f(x)\}^2 dx = \int \hat{f}_h^2(x)dx - 2E\{\hat{f}_h(x)\} + \int f^2(x)dx \tag{5}$$

The third term can be ignored since it does not depend on the bandwidth and (5) can be simplified as:

$$F(h) = \int \hat{f}_h^2(x)dx - 2E\{\hat{f}_h(x)\} \tag{6}$$

Various modifications of cross-validation methods have been proposed to accurately estimate the second term in (6).

(2) Plug-in methods

$$MISE(h) = \int E\{\hat{f}_h(x) - f(x)\}^2 dx = \frac{R(K)}{Nh} + \frac{h^4}{4}(u_2(K))^2 R(f'') + o(h^4) + o\left(\frac{1}{Nh}\right) \tag{7}$$

where

$$u_2(K) = \int u^2 K(u)du, \quad R(K) = \int K^2(u)du,$$
$$R(f'') = \int f''^2(u)du \tag{8}$$

Define the Asymptotic Mean Integrated Squared Error (AMISE) as:

$$AMISE(h) = \frac{R(K)}{Nh} + \frac{h^4}{4}(u_2(K))^2 R(f'') \tag{9}$$

So the asymptotically optimal bandwidth is:

$$h_{opt} = \left(\frac{R(K)}{u_2^2(K)R(f'')}\right)^{1/5} n^{-1/5} \tag{10}$$

In (10) only $R(f'')$ is unknown and has to be estimated. Thus, plug-in methods mainly concern the techniques to estimate $R(f'')$.

Among the two classes of methods, the most popular approach is Silverman's rule of thumb (ROT) [18] in which the density is regarded as the normal distribution, but it usually makes the results over-smooth in multimodal models. Further, it has been reported in [24] that one-sided cross-validation method (OSCV) [25] and Sheather–Jones plug-in method (SJPI) [26] are outstanding methods among various bandwidth selection methods. In this paper, the densities estimated by ROT, OSCV and SJPI are evaluated and compared with the results obtained by the proposed method. ROT and SJPI are introduced in [18] and [26], respectively. The calculation procedure for OSCV using Gaussian kernel is deduced in Appendix A.

2.2. Kernel density estimation via diffusion

The Gaussian kernel density estimator of (1) can be written in an alternative form:

$$\hat{f}(x;t) = \frac{1}{N}\sum_{i=1}^{N} \phi(x, X_i; t) \tag{11}$$

where

$$\phi(x, X_i; t) = \frac{1}{\sqrt{2\pi t}} \exp\left[-\frac{x - X_i}{2t}^2\right] \tag{12}$$

where $\sqrt{t}$ has the same definition as h in (1), referred to as the bandwidth.

It is shown in [27], the Gaussian kernel density estimator is the unique solution to the following diffusion partial differential equation (PDE):

$$\frac{\partial}{\partial t}\hat{f}(x;t) = \frac{1}{2}\frac{\partial^2}{\partial x^2}\hat{f}(x;t), \quad x \in X, \quad t > 0, \quad X \equiv R \tag{13}$$

with the initial condition of

$$\hat{f}(x;0) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - X_i) \tag{14}$$