Aeolian Research 16 (2015) 49–54

Contents lists available at ScienceDirect

Aeolian Research

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

Calculation of the aeolian sediment flux-density profile based on estimation of the kernel density

Meng Li *, Zhibao Dong, Zhengcai Zhang

Key Laboratory of Desert and Desertification, Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences, No. 322, West Donggang Road, Lanzhou, Gansu Province 730000, China

ARTICLE INFO

Article history:
Received 19 June 2014
Revised 11 November 2014
Accepted 11 November 2014
Available online 26 November 2014

Keywords:
Sediment flux-density profile
Nonparametric method
Kernel density estimation
Bandwidth

ABSTRACT

Aeolian sediment flux is an important issue of aeolian research. Parametric estimation is a traditional method in which aeolian sediment flux is estimated based on parameterization of a chosen equation. This method is simple, but has some limitations; specifically, it requires a priori assumptions about the density distribution that may not be correct. In this study, we applied a popular and extensively used, data-driven, nonparametric method called kernel-density estimation to calculate the aeolian sediment flux-density profile. Nonparametric methods make no prior assumption about the form of the density distribution to be estimated; instead, the aim is to obtain an empirical estimate from the data that can provably converge on the true density that would be obtained using an infinite sample size. Through the calculation of aeolian sediment flux based on kernel-density estimation, we determined that the key point in this method is not selection of the kernel function, but rather the selection of the optimal bandwidth, which is a difficult task. The results of our calculations showed that the method is both computationally feasible and acceptably accurate. Equally significantly, the idea of applying nonparametric methods to the calculation of aeolian sediment fluxes may lead to the development of a suite of other related analytical and modeling methods.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Wind erosion of soil leads to ecosystem degradation and various hazards to human values in arid and semiarid areas, which make up one-third of the world's surface (Lal, 1990; Sterk and Raats, 1996). This erosion damages valuable and nonrenewable soil resources, and the sediments generated in the process of erosion may form huge clouds that block sunlight, pollute water, damage crops and herds, and even threaten human life. In addition, climate and weather may be influenced by dust suspended in the atmosphere, since the dust reflects, scatters, diffuses, and absorbs solar radiation (Han et al., 2009). As human activities such as land reclamation and over-grazing can make the climate dryer and interact with any long-term warming and drying trends, the exposure of more soil to the wind exacerbates the problem of soil erosion (Dong et al., 2000). During the process of wind erosion, sediment particles are generated and transported by the wind in one of three modes (suspension, saltation, or creep), depending on the aerodynamic properties of the particles and the strength of the wind. Even within the same mode, particles vary in their speed, direction, acceleration, and other motion parameters. Due to these variations, particles are dispersed to different heights above the ground and form a sediment cloud.

Parametric estimation is a traditional method of aeolian sediment flux research that has been used to describe this cloud. In this approach, the aeolian sediment flux profile is assumed to be described by a mathematical function with several parameters. Some distribution functions have been widely adopted, such as the exponential and logarithmic distributions. After the distribution function has been chosen, its parameters are estimated according to the observed data. The literature on sediment flux research based on this approach includes data generated by wind-tunnel tests (Butterfield, 1999; Dong et al., 2006), field observations (Greeley et al., 1996; Namikas, 2003), and numerical simulations and theoretical analyses (Anderson and Haff, 1988, 1991; Zheng et al., 2004; Kang et al., 2008; Shi and Huang, 2010).

Compared with parametric estimation methods, nonparametric methods make no prior assumption about the form of the flux density to be estimated. They are therefore both flexible and capable of reducing modeling biases, and can potentially generate more robust and accurate estimates. Their biggest advantage is that a supposed distribution function is not required a priori, thereby avoiding the problem of inadvertently selecting an inappropriate

* Corresponding author. Tel.: +86 931 496 7485.
E-mail address: lmdasher@163.com (M. Li).

http://dx.doi.org/10.1016/j.aeolia.2014.11.003
1875-9637/© 2014 Elsevier B.V. All rights reserved.

model. In addition, where outliers exist in the data, parametric methods may fail to capture the complete structure of the actual curve. Non-parametric estimation methods alleviate this problem by treating each observation as a part of the model.

Non-parametric estimation methods include histogram estimation (Triola, 2010), Rosenblatt (1956) estimation, Parzen (1962) kernel-density estimation, and nearest-neighbor estimation (Wasserman, 2007). The histogram estimation method has been applied extensively because it is simple and intuitive, but the size range of the observed data must be known in advance, and the density estimation curve is discrete. For this reason, the Rosenblatt and Parzen kernel-density estimation methods were developed. Rosenblatt estimation does not require a subdivision strategy for the data and the intervals are calculated rather than assumed, so that data points always lie in the middle of the interval. It has been mathematically demonstrated that the estimator obtained is close to the true value (Rosenblatt, 1956). In Parzen kernel-density estimation, each estimated point has a fixed neighborhood. If the neighborhood size is large, dense data points exert excessive influence on the overall distribution, causing flattening of curves and potentially eliminating spikes that represent important information. In contrast, sparse points and outliers may be ignored because of their small neighborhood, and estimates for these neighborhoods may be zero even though a non-zero result would be more accurate or physically realistic. Loftsgaarden and Quesenberry (1965) developed nearest-neighbor estimation to mitigate the problems with Parzen kernel-density estimation.

In nonparametric methods, it is necessary to account for the bandwidth, which represents a smoothing factor that is used to reduce the effect of spikes in the density distribution (i.e., to account for the effect of outliers), thereby producing a more regular function that does not completely ignore the effects of outliers. When the bandwidth is large, kernel-density estimation functions better than the nearest-neighbor method, which is not recommended, but many scholars nonetheless use this method to sort the data. Efron (1979, 1982) and Efron and Stein (1981) presented a nonparametric estimation method called bootstrapping, which produced a model that fit the actual distribution, but with a relatively high error. Silverman and Young (1987) decreased the mean squared error (MSE) of the bootstrapping method. Katkovnik and Shmulevich (2002) proposed a variable-window kernel-density estimation method which requires only the knowledge of the variance of the estimate. By means of numerical simulations, this method performed significantly better than any constant-bandwidth method.

In this study, we examined the improved kernel-density estimation method developed by Parzen (1962) with the goal of identifying the key factors that affect the use of this method. We then applied the method to calculate the wind-blown sediment flux and compared the results with empirical data.

2. Kernel-density estimation

Kernel-density estimation attempts to estimate an unknown density function based on probability theory. This method has existed for decades and some early discussions on kernel-density estimations can be found in Rosenblatt (1956) and in Parzen (1962). Ruppert and Cline (1994) proposed a modified kernel-density estimation based on a clustering algorithm for the dataset's density function. As computers become more capable of handling high burden computation, research interests have increased.

2.1. The model definition

In our study, we started with the model of Parsen method. First, draw a random sample $X_1, X_2, \ldots, X_n$ from the density function $f(x)$. $K(\cdot)$ is a probability-density function for the kernel and n is the sample size. The kernel-density estimation for $f_n(x)$ is defined as:

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \quad \forall x \in R \tag{1}$$

where n represents the sample size, the positive constant h is called the bandwidth, i is the sample number, and R represents the set of real numbers.

2.2. Selection of kernel functions

Kernel-density estimation deals with more than just obtaining an appropriate sample; it also requires careful estimation of the kernel function and the bandwidth. All three factors determine the performance of the estimation. Kernel functions must meet the following requirements:

$$\text{Non-negativity} : K(x) > 0, \quad \forall x \in R \tag{2}$$

$$\text{Symmetry} : K(x) = K(-x), \quad \forall x \in R \tag{3}$$

$$\text{Normalization} : \int_{-\infty}^{+\infty} K(x) = 1 \tag{4}$$

Commonly used kernel functions (Wasserman, 2007) include the triangular, Epanechnikov, quartic, triweight, Gaussian, cosine, and exponential functions. In this work, we found that some kernel functions, such as cosine kernel, Epanechnikov kernel and quartic kernel functions, were not appropriate for the calculation of the aeolian sediment flux-density profile, because they are confined as $|(x - Xi)/h| \leqslant 1$. Only Gaussian kernel and exponential kernel are appropriate for this calculation. In practice, bandwidth selection becomes more important, as we will demonstrate in this paper.

2.3. Estimation of the bandwidth

It is important to choose an appropriate bandwidth to provide an accurate estimation of the kernel's density distribution. Ideally, the bandwidth should be as low as possible to avoid over-smoothing the curve, but high enough to remove spikes in the estimated distribution that would distort the description of the empirical data. In a univariate case, the performance of the kernel-density estimation depends strongly on the bandwidth, which functions as a weight function for the estimated kernel. Selection of the optimal bandwidth is a crucial problem in kernel-density estimation and has been the subject of considerable theoretical research, especially in the context of univariate kernel-density estimation (Dutta, 2011). These efforts include studies by Rudemo (1982), Bowman (1984), Silverman (1986), Scott and Terrell (1987), Park and Marron (1990), Jones and Kappenman (1991), Cao et al. (1994), Marron and Ruppert (1994), Wand and Jones (1995), and Simonoff (1996).

The method for global and local bandwidth selection is the mean integrated square error (MISE) criterion (Wasserman, 2007). Familiarization with the MISE criterion is not required for the practical use of the kernel-density estimation, but it will help those who are interested, learn how one rigorously arrives to a well-chosen bandwidth.

$$\text{MISE} = E\left\{\int \left[\hat{f}(x) - f(x)\right]^2 dx\right\} = \int E\left[\hat{f}(x) - f(x)\right]^2 dx \tag{5}$$

where the density estimation $\hat{f}(x)$ is a kernel-density estimation of $f(x)$ and is a function of the bandwidth h, E represents the statistical expectation. Our approach of bandwidth estimation is to minimize

Download English Version:

https://daneshyari.com/en/article/6426318

Download Persian Version:

https://daneshyari.com/article/6426318

Daneshyari.com