The Kernel Estimation in Biosystems Engineering

Esperanza Ayuga Téllez
Departamento de Economía y Gestión Forestal. E. T. S. I. de Montes.
Universidad Politécnica de Madrid.
28040 Madrid. Spain

Mª Ángeles Grande Ortiz
Departamento de Física y Mecánica Aplicadas a la Ingeniería Agroforestal. E. T. S. I. Montes.
Universidad Politécnica de Madrid.
28040 Madrid. Spain

Concepción González García
Departamento de Economía y Gestión Forestal. E. T. S. I. Montes.
Universidad Politécnica de Madrid.
28040 Madrid. Spain

Ángel Julián Martín Fernández
Departamento de Matemática Aplicada a los Recursos Naturales. E. T. S. I. Montes.
Universidad Politécnica de Madrid.
28040 Madrid. Spain

Ana Isabel García García
Departamento de Planificación y Proyectos. E. T. S. I. Agrónomos.
Universidad Politécnica de Madrid.
28040 Madrid. Spain

ABSTRACT

In many fields of biosystems engineering, it is common to find works in which statistical information is analysed that violates the basic hypotheses necessary for the conventional forecasting methods. For those situations, it is necessary to find alternative methods that allow the statistical analysis considering those infringements.

Non-parametric function estimation includes methods that fit a target function locally, using data from a small neighbourhood of the point. Weak assumptions, such as continuity and differentiability of the target function, are rather used than “a priori” assumption of the global target function shape (e.g., linear or quadratic).

In this paper a few basic rules of decision are enunciated, for the application of the non-parametric estimation method. These statistical rules set up the first step to build an interface user-method for the consistent application of kernel estimation for not expert users. To reach this aim, univariate and multivariate estimation methods and density function were analysed, as well as regression estimators. In some cases the models to be applied in different situations, based on simulations, were defined.

Different biosystems engineering applications of the kernel estimation are also analysed in this review.

Key words: Non-parametric estimation, kernel methods, simulation, biosystems engineering.

1. INTRODUCTION

Statistical information from different fields of biosystems engineering usually violates assumptions necessary to analyse that information with traditional (parametric) methods. Nonparametric estimation provide an alternative series of statistical techniques that require no or very limited assumptions to be made about the data.

There is a wide range of nonparametric methods that can be used in different circumstances, among them we find the non-parametric estimation of probability density functions which includes methods that fit a target function locally, using data from a small neighbourhood of the point. Weak assumptions, such as continuity and differentiability of the target function, are rather used than “a priori” assumption of the global target function shape (e.g., linear or quadratic) [1].

In this way, non-parametric estimation techniques has been improved thanks to the computers development, which has allowed to use some procedures that has been suggested before. The first works refered to these techniques are published between 1930 and 1950. Nevertheless, till the eightieth decade, it have not been applied to real data. During ninetieth years publications about theoretical studies of non-parametric estimators have been numerous and some papers with applications to real data appeared commonly. Actually are increasing many applications in economy field.

It is remarkable that the increasing use of these techniques is not due to a unique circumstance. The novelty, applicability and theoretical properties give the method valid results whit unquestionable interest. For this reason, is necessary statistical techniques reviewed and complete study of their possible applications, focused on the biosystems engineering most used.

2. ESTIMATION OF THE DENSITY FUNCTION

The non-parametric estimation of the density function most studied mathematically and those that have a higher number of applications to real data are those that are based on the definition of a Kernel function [2, 3, 4]. To use these it is necessary to choose a smoothing parameter as well as the kernel function. Both will determine the final expresión of the estimate of the density function.

The non-parametric estimator, Eq. (1), starts from kernel function K(x).

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \qquad (1)$$

were h is the smoothing parameter and $X_1...X_n$ the recorded data. The value of h is a positive number that is calculated to minimize some type of error

Nadaraya [4] establishes the following properties for K(x).

a) $K(x) = K(-x)$

b) $\int K(x)dx = 1$

c) $\sup\limits_{-\infty < x < \infty} | K(x) | \leq A < \infty$

d) $\int x^i K(x)dx = 0, \quad i = \overline{1,\ s-1}$

with s even and greater o equal to 2;

e) $\int x^s K(x)dx \neq 0$

f) $\int x^s | K(x) | dx < \infty$

It has been proved that, under certain conditions of regularity of K and of f and if $h \to 0$ when $n \to \infty$, the kernel estimator is asymptotical biased and normally distributed.

Martínez-Falero et al. [5] and Ayuga et al. [6] establish the methodology to select the kernel function and the procedure to estimate the smoothed parameter more appropriate, depending on the sample characteristics (variance, skewness, kurtosis and number of modes). The election of the kernel function change according to the characteristics of the information but it is recommended the maximum likelihood cross-validation method [7 and 8].

The applications of this methodology are mainly in the fields of research and forest ecosystems management. For example, kernel methods were employed to a statistical canopy reconstruction to infer the geometrical features of a broad-leaf tree crown from a foliage sub-set (leaf data sample) [9]; to fit bimodal diameter distributions which are typical of temperate forest populations with several species or populations under extensive forest management [10]. In another way there is modified the management of some pine-groves to make them more sustainable in [11 and 12] with the application of the kernel method to statistical distributions of forest populations variables.

Allowing variety it can be found some water resources applications: with historical floods and paleoflood information [13]; Moon and Lall [14] considered the direct kernel estimation (KQ) of the quantile function with special boundary kernels to extrapolate beyond the largest recorded flow. Vogel and Fennessey [15] investigated nonparametric estimation of annual flow duration curves using daily flow data.

To simplify the process of cost estimation in many engineering applications, deterministic cost models are often used. Single value cost estimates are invariably the output of these models. It has been recognized by the cost estimating community that estimates should more appropriately be given in terms of probability distribution functions (PDFs) and that simulation methods are superior to analytical methods for determining these cost distributions. A simulation approach employing non-parametric estimation techniques and their asymptotic properties in the development of the PDFs of cost estimates is proposed to the problem of project bidding [16].

3. NON PARAMETRIC REGRESSION

When two or more variables are studied jointly, it is logical to try to ascertain and evaluate the degree of relationship among them; however, when seeking a functional relationship among variables, one must search for a means to express the form of this relationship.

The use of parametric regression involves a series of restrictions on the general nature of the problem -the prior hypotheses that have to be fulfilled, which are frequently far from reality, and the specific functions to be fitted-. Thus the general equations obtained may hide important features of the relationship.

So nonparametric regression methods are more advisable to use in problems whose data are suspected not to fulfil the requirements of classical methods. On the other hand, nonparametric estimation of the regression cannot provide a general simple expression of that relationship. It is necessary to consider also the difficulty that the application of nonparametric methods has for researchers.

The methodology proposed in [17] and [18] tries to introduce and facilitate application of nonparametric regression by researchers. On the one hand, the choice of a suitable kernel [19] and bandwidth estimators [20, 21] is pointed out. On the other, the problem of the tendency of the regression curve to drop sharply at the edges of the interval containing the sample data is solved by the use of a pseudosample. The procedure involves a clear improvement in the fit of the regression curve to the sample data. The simplicity and general nature of the procedure used make it easily applicable to any type of data.

In [17] study by means of nonparametric curves is more effective than by the use of other classical statistical techniques, in comparing the different antiinflammatory activity of several agents and with the relation to elapsed time from administration of carragenine.

Due to the difficulty of correctly using the parametric regression methods for the study of dasometric variables (including big data volumes) Ayuga et al. [18] consider the methods of kernel regression to be much more suitable to model the relations among the variables.]. In the study of Pinus sylvestris woodland in Cercedilla, this method can capture relationships that are not detected by the usual parametric model. The main conclusions of this paper are that a linear relationship between heights and crown exists, and height and trunk diameter. In addition, there is a non-linear and monotonous increasing relation, between the area and crown volume, opposite to the trunk diameter. Finally, this study allows us to estimate relationships between the orientation and the diameter of the trunk and between the ground's slope and the tree's height. Also a relationship is observed between height and crown area that parametric methods do not show, and this relationship reveals interesting variations.

In a study [22] based on the employment of several statistical models, included the regression by means of kernel functions, applies to the biometric information obtained a breeding farm of red partridge (*Alectoris rufa sp.)*, it is proved that is the most suitable method to assign the age to the young birds. To adjust the non-parametric model, diverse variables have been used, and the better relation for the objective is the measurement between the chick age and the length of the beak height.

Bradley and Potter [23] describe a new approach for flood frequency analysis (FFA) of model-simulated flows by smoothed peak discharge vs 3-day flow volume for regulated and unregulated conditions, en route to developing a Peak to Volume FFA that may be useful for examining the impact of flow regulation on floods.

Baier and Cohn [24] smooth atmospheric concentrations of selected constituents versus precipitation, to remove the effect of precipitation variability on acid deposition trends.

Lall and Bosworth [25] look for relationships between precipitation, evaporation, net precipitation and annual inflow into the Great Salt Lake.

The number of vehicle miles traveled (VMT) each year by households is a key variable of interest [26]. It is used in most models of travel demand, as a control variable, a response variable, or both. This study employs nonparametric econometric techniques to examine the effects of household income, vehicle ownership and workers on annual household VMT using the 1995 Nationwide Personal Transportation Survey. The results are density functions and regression surfaces for VMT, in relation to these and other variables, including public transit availability, housing location (urban versus rural), and retirement.

Mudelsee et al. [27] review the reconstruction and assess the data quality of the records, which are based on combining documentary data from the interval up to 1850 and measurements thereafter, finding both the Elbe and Oder records to provide reliable information on heavy floods at least since A.D. 1500. They explain that the statistical method of kernel occurrence rate estimation can overcome deficiencies of techniques previously used to investigate trends in the occurrence of climatic extremes. The observed trends are shown to be both robust against data uncertainties and only slightly sensitive to land use changes or river engineering, lending support for climatic influences on flood occurrence rate. They finally draw conclusions about flood disaster management and modeling of flood occurrence under a changed climate.

Extreme values of weather and climate variables sit per definition at the tails of the probability density function (PDF). The tail probability is for most variables rather small: extreme weather events are rare. Mudelsee [28] reviews methods to estimate the time-dependent occurrence rate, $\lambda(t)$, using univariate observations. So two implementations of the continuous-time peak-over-threshold (POT) technique are presented: parametric and nonparametric. The non-parametric implementation (kernel smoothing) uses continuously shifted time intervals to explore the time dependence. Kernel smoothing allows to construct confidence bands for $\lambda(t)$ by means of bootstrap resampling. The kernel estimation has no parametric restrictions, it allows also non-linear and non-monotonic $\lambda(t)$. It illustrates kernel occurrence rate estimation with two examples: (1) the risk of extreme floods of the Elbe river over the past 1000 years and (2) the risk of major windstorms in the North Sea region over the past 50 years.

4. SPATIAL ANALYSIS

To this sort of studies it is suggested the method described by Silverman [29] validated for bivariate density functions estimating with greater sample sizes than 70. A direct application is the estimation of the spectral density function, using Cacoullos's procedure [30] and Nadaraya [4], checked by other authors for bivariates spectral densities.

The estimation of the function of spectral density has been applied to the study of the forest structures [31, 32].

Kriging is the most popular method for spatial interpolation in the Earth Sciences. It is a clever, constrained, parametric regressor that is sometimes called nonparametric. Yakowitz and Szidarovsky [33] developed theoretical results to establish the consistency and convergence properties of Kriging. For Kriging to work, they showed that proper variogram selection was critical. They formalized a Nadaraya Watson kernel regression estimator for spatial regression, established its consistency and convergence rates, utility for estimating functionals (e.g., integrals or derivatives of the surface), and developed a nearest neighbor estimator of the local mean square error of estimation.

Lall and Ali [34] consider the recovery of subsurface stratigraphy from drill log information. This information is encoded in a binary function (1 for ``sand,'' 0 for ``clay'') and two models to interpret this data are developed. The first considers the occurrence of sand in the vertical as a nonhomogeneous Poisson process (NPP) at any point in the domain and uses a 3-dimensional kernel method to interpolate this rate from neighboring drill logs.

The method of kernel estimation has been used to develop spatially continuous seismicity models (earthquake probability distributions) from a given earthquake catalogue [35, 36].

Other applications of the kernel estimation for multivariate density function, will be the following ones:

Fire and weather archive data for the province of Ontario and Canada were investigated using spatial statistical and time series analysis methodologies [37]. Spatial point pattern analysis was used to investigate spatial patterns of lightning-caused fire occurrence in Ontario. Lightning-caused forest fires were found to be spatially clustered. Evidence was found that this clustering follows the spatial pattern of lightning strikes on dry weather days. Kernel estimation of the spatial intensity of lightning strikes on days when an element of the Fire Weather Index (FWI), the Duff Moisture Code (DMC), exceeded 20 provided a spatial pattern quite similar to that of lightning-caused fires. Localized dry weather and lightning storm occurrence are the principal determinants of the spatial clustering of lightning caused fire occurrence.

Sabel et al. [38] outlines the development of a method for using kernel estimation cluster analysis techniques to automatically identify road traffic accident ‘black spots’ and black areas’. Christchurch, New Zealand, was selected as the study area and data from the Land Transport New Zealand crash database used to trial the technique. Point analysis using kernel estimation is integrated functionality in modern GIS packages, such as ESRI’s ArcGIS, however determining the statistically significance is not. So a GIS and Python scripting was used to implement the solution, combining spatial data for average traffic flows with the recorded accident locations. Kernel estimation was able to quickly identify the accident clusters, and when used in conjunction with Monte Carlo simulation techniques, was able to identify statistically significant clusters.

Stanley [39] presents an overview of the perspectives and techniques related to system identification. The author uses those techniques within the context of several well-defined problems in the nervous system. Also it is remarqued that several of the problems presented involved combinations of non-parametric and parametric representations. It is emphasized the utility of nonparametric techniques to the analysis of complex systems, where little is known about the underlying dynamics.

Sigalotti et al. [40] report a method that converts standard smoothed particle hydrodynamics (SPH) into a working shock-capturing scheme without relying on solutions to the Riemann problem. Unlike existing adaptive SPH simulations, the scheme presented is based on an adaptive kernel estimation of the density, which combines intrinsic features of both the kernel and nearest neighbor approaches in a way that the amount of smoothing required in low-density regions is effectively controlled.

5. CONCLUSIONS

A kernel estimation method needs easier requirements. The rules described in this work, allow unifying the criteria for the application of these kinds of methods to estimate univariate density functions, regression models and spatial analysis.

Its applications in biosystems engineering are increasingly numerous. Specifically, is widely used in the forest and water resources management sciences. Also, it is necessary to emphasize its role in natural risks forecasting.

6. REFERENCES

[1] U. Lall, “Recent advances in nonparametric function estimation: Hydrologic applications. U.S. National Report to IUGG, 1991-1994” **Rev. Geophys**. Vol. 33 Suppl., American Geophysical Union. 1995.

[2] M. Rosenblatt "Remarks on some non-parametrics estimates of a density function". **Annl. Math. Stat**., Vol 27, 1956. pp 832-837.

[3] E. Parzen, "On estimation of a probability density function and mode". **Annl. Math. Stat**., Vol. 33, 1962. pp 1065-1076.

[4] E. A. Nadaraya, **Nonparametric estimation of probability densities and regression curves**. Kluwer Academic Publishers, Dordrecht. 1989.

[5] J. E. Martínez-Falero, E. Ayuga, C. Gonzalez. Estudio comparativo de distintas funciones núcleo para la obtención del mejor ajuste según el tipo de datos. **QUESTIIÓ**, Vol. 16, 1,2 y 3,. 1992. pp. 3-26.

[6] E. Ayuga, C. González, J. E. Martínez-Falero, “Elección de la función núcleo y ancho de banda más apropiado para la estimación de funciones de densidad con muestras grandes”. **XXI Congreso Nacional de Estadística e Investigación Operativa.** 1994. pp. 332-333.

[7] R. P. Duin, "On the choice of smoothing parameters for Parzen estimators of probability density functions". **IEEE Transactions on Computers**, C-25, 1976. pp. 1175-1179.

[8] J. Hermans and J. D. F. Habbema. **Manual for the ALLOC discriminant analysis programs**. Universidad de Leiden, Dept. de Estadística Médica. 1976.

[9] R. Giuliania, E. Magnaninia, F. Nerozzib, E. Muzzia and H. Sinoquetc. “Canopy probabilistic reconstruction inferred from Monte Carlo point-intercept leaf sampling”, **Agricultural and Forest Meteorology**. Vol. 128, 1-2, 20. 2005, pp 17-32.

[10] J. M. Torres-Rojo. “Prediction of multimodal diameter distributions through mixtures of Weibull distributions”. **Agrociencia**. Vol. 39. 2005, pp 211-220.

[11] A. García; P. Irastorza; J. García; E. Martínez Falero; J. Solana and E. Ayuga. “Concepts associated with deriving the balanced distribution of uneven-aged structure from even-aged yield tables: application to Pinus sylvestris L. in the central mountains of Spain”, in **Management of mixed-species forest: silviculture and economics**. IBN-DLO, Wageningen , 1999. pp. 108-127.

[12] E. Ayuga, E.; A. García; M. A. Grande; S. Martín and P. A. Tiscar. “Curvas orientativas de la distribución equilibrada de Pinus nigra Arn. a partir de las tablas de producción de masa regular para el Sistema Ibérico” in **Los Pinares de Pinus nigra Arn en España: Ecología, Uso y Gestión**, FUCOVASA, Madrid, 2005. pp. . 343-368.

[13] S. L. Guo, “Nonparametric Variable Kernel Estimation With Historical Floods and Paleoflood Information”, **Water Resour. Res.** Vol. 27(1), 1991. pp. 91–98.

[14] Y. Moon and U. Lall, “A Kernel Quantile Function Estimator for Flood Frequency Analysis”, **Water Resour. Res.**, Vol. 30 (11), 1994. pp 3095-3103.

[15] R. M. Vogel and N.M. Fennessey, “Flow Duration Curves I.: New Interpretation and Confidence Intervals”, **J. Water Res. Plng. Mgmt.**, Vol. 120 (4), 1994. pp 485-504.

[16] Y. Asiedu and R. W. Besant. “Simulation-based cost estimation under economic uncertainty using kernel estimators”. **International Journal of Production Research.** Volume 38, Number 9. 2000. pp 2023 – 2035.

[17] E. Ayuga, C. Ayuga, C. González, J.E. Martinez Falero, “Estimation of non-parametric regression in the analysis of the anti-inflammatory activity of diverse extracts of Sideritis foetens”. **Journal of Applied Statistics**, Vol. 24, nº 5. 1997. pp. 559 - 572

[18] E. Ayuga, A. Martín, C. González, J.E. Martinez Falero, “Estimation of non-parametric regression for dasometric measures”. **Journal of Applied Statistics**. Volume 33, Number 8. 2006. pp. 819 - 836

[19] E. Ayuga. **Modelos no paramétricos de ajuste de curvas aplicados al ámbito forestal**. Tesis doctoral. Universidad Politécnica de Madrid. 1992

[20] T. Gasser, A. Kneip, and W. Köhler, W. "A Flexible a fast Method for Automatic Smoothing". **Journal of the American Statistical Association**, Vol. 86, 1991. pp 643-652.

[21] H. G. Müller, **Nonparametric regression analysis of longitudinal data**. Springer-Verlag, Berlín. 1988.

[22] E. Ayuga, M.J. Rueda and J.E. Martinez-Falero. “Estudio Biométrico de Pollos de Perdiz Roja (Alectoris Rufa L.) durante las tres primeras semanas”. Montes. Vol. 37. 1994. pp. 52-56.

[23] A. A. Bradley and K.W. Potter, “Flood Frequency Analysis of Simulated Flows”, **Water Res. Res**., Vol. 28 (9), 1992. pp. 2375-2386.

[24] W. G. Baier and T.A. Cohn, “Trend Analysis of Sulfate, Nitrate and pH Data Collected at National Atmospheric Deposition Program/National Trends Network Stations Between 1980 and 1991”, **U.S.G.S**., Reston, VA, 1993.

[25] U. Lall and K. Bosworth, “Multivariate Kernel Estimation of Functions of Space and Time Hydrologic Data, in Stochastic and Statistical Methods” in **Hydrology and Environmental Engineering**, edited by K. Hipel, Kluwer, Waterloo, 1993.

[26] Y. Kweon, and K. Kockelman, “Nonparametric Regression Estimation of Household VMT”. **2004 Annual Meeting of the Transportation Research Board.**

[27] M. Mudelsee, M. Börngen, G. Tetzlaff, and U. Grünewald, “Extreme floods in central Europe over the past 500 years: Role of cyclone pathway “Zugstrasse Vb””, **J. Geophys. Res.**, Vol. 109. 2004. pp 10-29.

[28] M. Mudelsee, “Weather extremes and global change: estimating time-dependent climate risk”, **Geophysical Research Abstracts**, Vol. 7, 2005.

[29] B. W. Silverman, **Density estimation for statistics and data analysis**, Chapman & Hall, New York, 1986.

[30] T. Cacoullos, "Estimation of a multivariate density". **Ann. Inst. Stat. Mathc**, Vol. 18,. 1966. pp. 178-189.

[31] A. Martín.; J.E. Martínez-Falero; S. Martín; E. Ayuga; A. García; C. González; M. A. Grande, “Aplicación del análisis espectral a la modelización de la distribución y la correlación espacial entre variables forestales”. **Proceedings of IX Congreso del Grupo de Métodos Cuantitativos, Sistemas de Información Geográfica y Teledetección. Tecnologías Geográficas para el desarrollo sostenible.** Edas. I. Aguado y M. Gómez, Universidad de Alcalá. 2000. pp 1-15.

[32] A. Martín.; S. Martín; C. González, E. Ayuga. “Aplicación del análisis espectral al estudio de las masas forestales”. Proceedings of **III Congreso Forestal Español**. Consejería de Medio Ambiente. Junta de Andalucía. Vol II, 2001. pp. 29-35.

[33] S. J. Yakowitz and F. Szidarovsky, “A comparison of Kriging with nonparametric regression methods”, **J. Mult. Anal**., 16 (1), 1985. pp. 21-53.

[34] U. Lall, and A.I. Ali, “Nonparametric Stratigraphic Interpretation from Drill Log Data”, Utah Water Res. Lab., **Utah State Univ**., 1992.

[35] C. Stock and E. G. C. Smith, “Adaptive kernel estimation and continuous probability representation of historical earthquake catalogues”. **Bull. Seism. Soc. Am**. Vol. 92. 2002. pp 904-912.

[36] C. Stock and E. G. C. Smith. “Comparison between seismicity models generated by different kernel estimations”. **Bull. Seism. Soc. Am**. 92. 2002. pp. 913-922.

[37] J. Podur. **Spatial and Temporal Patterns of Forest Fire Activity in Canada**. Degree Thesis. University of Toronto. 2001.

[38] C. E. Sabel, S. Kingham, A. Nicholson, Phil Bartie. 2005. Road Traffic Accident Simulation Modelling - A Kernel Estimation Approach. **17th Annual Colloquium of the Spatial Information** Research Centre University of Otago, Dunedin, New Zealand.

[39] G. B. Stanley, “Neural System Identification”, in **Neural Engineering**, edited by Bin He, Springer. 2005.

[40] L. Sigalotti, H. López, A. Donoso, E. Sira, J. Klapp. “A shock-capturing SPH scheme based on adaptive kernel estimation”. **Journal of Computational Physics**. Vol. 212(1): 2006. pp. 124-149.