VISUAL BASIC APPLICATIONS AND SPREADSHEET FOR TEACHING ESTIMATION OF NONPARAMETRIC DENSITY AND REGRESSION FUNCTIONS ®

Federico Palacios-Gonzalez
University of Granada
Spain

The aim of this paper is to show how teachers can use Visual Basic language on the spreadsheet for student sin order to gain the skills need in using nonparametric techniques for density and regression function estimations. The author has built a useful didactic tool on Excel Visual Basic for teaching and practice of nonparametric kernel methods, being intended for students with some preliminary knowledge on this topic. This Visual Basic Application (VBA) is loaded into Excel as a MACRO (or into the modules of a Workbook for EXCEL). The specific user functions incorporated into it are easy tools for students to obtain an intuitive perception of nonparametric estimation for density and regression functions. The VBA also allows Excel to make use of an added menu similar to a small Statistical Package specialised in nonparametric methods.

INTRODUCTION.

Statistical packages are professional tools aimed at experts but not students. Results from data can be obtained efficiently, but processes of calculation are not shown; concepts are usually introduced at different levels and even with different terminology from the one employed in the teaching class. Besides this, methods and concepts are frequently mixed with others unnecessary for our teaching purposes and, usually, our teaching topics are only a fraction of the professional package content. For these reasons, there is need of a new didactic tool between the theory and the professional statistical packages.

A specific didactic package written in VB for Excel and specialized in teaching a particular topic has the following advantages. All the steps of the calculus process can be shown in the same order and using the same notation and terminology as the one given in the theoretical sessions. The didactic package can be oriented to show the behavior and characteristics of the statistical methodology being taught. It can be organized in a similar way to a professional package working as a bridge between the theoretical knowledge taught and the examples practiced to get an optimal comprehension of the topic. It can also be combined with simulation techniques to observe its efficiency in controlled situations where the solutions are previously known. It can make use of all the didactic spreadsheet advantages, in particular those from Excel, leaving out the inconveniences. Wide commentary and examples of the advantages and disadvantages of the didactic spreadsheet uses, when teaching statistics, can be seen in Hunt (1994), Hall (1995), Hunt (1996), Nash and Quon (1996), Gatti and Harvell (1998), Horgan (1999), Beauchamp and Youssef (1998), Bell (2000), Evans (2000).

The present paper describes the didactic use of *NonParametric.xla*, which is a MACRO, containing a set of user functions and a specific package to teach nonparametric density estimation and nonparametric curve estimation using kernel methodology. This tool is developed in Visual Basic language for Excel and makes use of the user-friendly interactivity and graphic capability of the spreadsheet.

THE USER FUNCTIONS FOR NONPARAMETRIC METHODOLOGY.

Here we show the didactic use of *user functions* whose code is, in the MACRO, named *NonParametric.xla*, or, alternatively, in the Excel workbook, named *NonParametric.xls*. The user functions for Kernel Density Estimation (KDE) are *EpanechDensity*, *GaussianDensity* and *KernelDensity*; these functions constitute the initial tools for showing the behavior of this methodology in Excel and use the following estimator by Rosenblatt (cited in Härdle, 1991:45)

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

EpanechDensity uses Epanechnicov kernel, *GaussianDensity* uses Gaussian kernel and *KernelDensity* can use any of the seven kernels listed in Härdle (1991:45). In a similar way, the three regression functions: *EpanechRegression*, *GaussianRegression* and *KernelRegression* use the Nadaraya-Watson estimator

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)}. \qquad (2)$$

By now students have theoretical knowledge on the topic, but they have no intuitive perception of the methodology behavior and no experience in using it either. Using the functions above students can obtain an intuitive knowledge of the following effects: the effect of bandwidth, the effect of different kernels, the effect of the sample size and the effect of estimate variability due to sample randomness. At the beginning these effects will be shown one by one independently and then, in successive steps, all together, in a compound form. All the user functions above can be used as any habitual implemented function of Excel from the *Function Manager Assistant*.

In order to illustrate how this tool can be managed, we shall use a known Gamma density model to simulate several samples with several sizes so that the different effects mentioned above be exhibited. In particular

$$f(x;\alpha = 2,\theta = 10) = \frac{1}{100} x\, e^{-x/10}. \qquad (3)$$

Any other model can be used to produce similar and many more examples including classical sets of real data too, such as Old Faithful geyser (Loader, 1999:421 and 424) in the training of students.

In a similar way the following model for the study of the behavior of kernel curve estimation can be used with ε_x being uncorrelated and with normal distribution $N(0,1)$

$$y = ae^{bx} + c\cos\left(\frac{2\pi x}{T}\right) + \sigma\varepsilon_x. \qquad (4)$$

Several values of parameters can be taken in order to explore their influence on kernel method effectiveness. Any other model (including heteroscedasticity and autocorrelation) or real data can also be used.

EXPLORING THE FOUR EFFECTS ABOVE

From an Excel chart we can compare the real density and the estimated densities under different conditions. Such a comparison between the model and the estimations gives the students a clear perception of the methodology behavior. Figure1 is a template that is the final result of a constructive process in four steps: in the first step, a simulated sample from density (1) of size n= 36 is allocated in column A, in column B there is a grid of x values, column C contains the true density (1) valued on the column B grid, and column D contains the estimated density on the grid using the sample in column A and the *EpanehDensity* function. The bandwidth is allocated into cell F3. Changing the bandwidth values into F3, the column D is recalculated and shaped into the template chart that is adapted to the new results. In an interactive way students can perceive the effect of the bandwidth on the smoothness and bias of the estimate density. Any other sample size can be used if further practice is required.

In the second step, the template given in Figure 1 is extended to show the students similar results for several kernels. The respective bandwidths must be selected in a convenient way. In this case Column E contains the estimated density but using the *GaussianDensity* function. For each Gaussian bandwidth proposed in F5 the respective Epanechnicov bandwidth is calculated by the Canonical Bandwidth Transformation (Härdle, 1991:74) is implemented into the *ConvertBandwidth* function allocated into cell F3. Changing the bandwidth in F5, all the columns, and also the chart, fit the new situation. The interactive handling of such a template is an interesting simulated experience comparing the similarity results obtained with different kernels

and their bandwidth effects. The behavior of other kernels can also be compared using *KernelDensity* function.

In the third step, the template of the same Figure 1 can also be adapted to explore the sample size effect: augmenting the sample size in column A (to n = 200, for example), taking equal kernels and bandwidths in columns D and E, but a reduced range of data (A3:A32 for example) in column D and the complete range in column E. The different smoothness degree due to the sample size can be appreciated. Different bandwidth values for kernels in combination with different sample size structures can be essayed to exploit this template conveniently.

In the fourth step the randomness effect in combination with the sample size and the bandwidth selection can be observed by changing column A. Now the sample is not fixed but volatile (any time that key F9 is pressed down). To reach such it is necessary to introduce the function "=GAMMAINV(RANDOM(); 2; 10)" into the cells of column A

Following the same process we can obtain a similar template for the curve estimation using the respective *EpanechRegression*, *GaussianRegression*, and *KernelRegression* functions. See Figure 1.

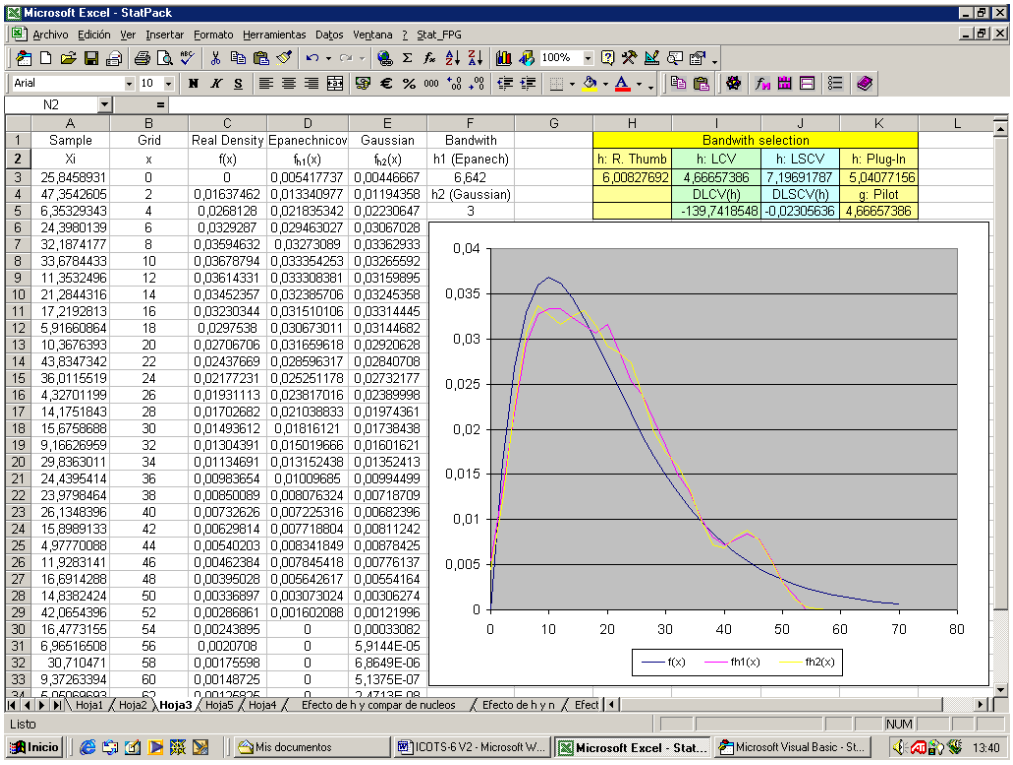

Figure 1. Curve Estimation

BANDWITH SELECTION.

At this point the students have a clear knowledge of the importance of selecting the right bandwidth but it is necessary to provide them with the didactic tool to do it. A great effort is being made to find an effective methodology to select the smoothing parameter correctly and quickly. With a certain periodicity reviews and comparative studies of known methodologies appear in the literature. This can be seen in Park and Marron (1990), Park and Turlach (1992), Sheater (1992), Cao et al. (1994), Jones et al. (1996), Farmen and Marron (1999), Loader (1999). It is convenient to remark that with respect to the global bandwidth selection methods, *Plug-In* seems does not seem to have more advantages than the *classical methods* as it was initially believed:

> *We find the evidence for superior performance of plug-in approaches is far less compelling than previously claimed……. We do not claim that the classical approaches to bandwidth selection such as AIC and cross validation will always produce the best estimates, but rather that, used properly, the results will often be far more*

informative than other recent work in bandwidth selection suggests (Loader, 1999:416,435).

The local bandwidth selection methods seem more effective than the global methods but no total agreement exists among authors referring to that:

The mean integrated squared error (MISE) of the adaptative methods is compared of a well-respected constant bandwidth often referred to as the Sheather-Jones plug-in (SJPI). A surprising fact is that an alternative visual error criterion is that the MISE performance of the SJPI is often quite close to that the adaptive methods (Farmen & Marron, 1999:143).

Taking into account the didactic objective of the present paper, we will concentrate on the global bandwidth selection. We will use the following user's functions: *DRuleOfThumb*, *DLCV* (Likelihood Cross Validation for Densities), *DLSCV* (Least Squared Cross Validation for Densities) for kernel density, and *DplugIn*, *RruleOfThumb*, *RCV* (Least Squared Cross Validation) and *GCV* (Generalized Cross Validation) for kernel regression. It is understood that the students are now familiar with the concepts involved. The bandwidth optimizing the cross-validations functions is searched into the convenient range. In a first step it is done, using the graphs of these functions and, in a second one, using the optimizer named *Solver* already implemented into the spreadsheet. Figure 1 shows an example of the template (H1:K5) using the functions for the density estimation case.

Rules of thumb and Plug-In bandwidth selection methods for kernel density estimation implemented in the user functions are based on the Mean Integrated Squared Error (MISE)

$$MISE\left(\hat{f}_h\right) = \int_{-\infty}^{+\infty} E\left[\left(\hat{f}_h(x) - f(x)\right)^2\right] dx \,. \tag{5}$$

Taylor expansion concludes in the following approximation (Hardle, 1991)

$$MISE\left(\hat{f}_h\right) = \frac{1}{nh} R(K) + \frac{h^4}{4}\left(\mu_2(K)\right)^2 R(f'') + o\left((nh)^{-1}\right) + o\left(h^4\right) \tag{6}$$

where f'' is the second derivate of the unknown target density,

$$R(K) = \int_{-\infty}^{+\infty} K(u)^2\, du \qquad R(f'') = \int_{-\infty}^{+\infty} f''(x)^2\, dx \quad \text{and} \quad \mu_2(K) = \int_{-\infty}^{+\infty} u^2 K(u) du \,. \tag{7}$$

The Asymptotic MISE optimal bandwidth is

$$h_0 = \left(\frac{R(K)}{R(f'')\left(\mu_2(K)\right)^2 n}\right)^{1/5} . \tag{8}$$

Unfortunately $R(f'')$ factor in (8) is unknown. The strategy followed to obviate this unknown factor, determine a rule of thumb or any of several Plug-In Methods (for more details see Härdle 1991, and Loader 1999).

Methods of Likelihood Cross Validation (LCV) for density estimation, Generalized Cross Validation (GCV) for curve estimation and Least Squared Cross Validation (LSCV) for both are very intuitive and of easy comprehension for students; details of these topics and their theoretical treatment can be found in Silverman (1986), Härdle (1991) and Eubank (1988).

EXPLOITING THE SPECIFIC PACKAGE.

As has already been said, a specific package with nonparametric methodology has been implemented into Excel. It is time to observe the concepts (already manipulated with the user's functions) from the package. The PopUp menu bars *ICOTS-6: Nonparametric* reproduces the appearance of a professional one. Most of the advantages of the statistical packages are incorporated into it or can be incorporated progressively by the author. Most of the professional package inconveniences (for example, the opacity in the processes of calculations) can be obviated. Both the possibility of adaptation from one year to the next one and the harmonization with the teaching strategy at any moment is possible. The advantages of the spreadsheet are always present and the limitations of the spreadsheets are overcome building appropriated Visual Basic Complements.

Figures 2 and 3 show details of the PopUp menu bars *ICOTS-6: Nonparametric* containing the procedures of package. This menu is allocated into the principal menu bar of Excel and it is available from any active worksheet at any time the work session as soon as *NonParametric.xla* is loaded into the spreadsheet. Optionally the *NonParametric.xla* can be installed in Excel as a permanent complement. Input and output data are handled into the spreadsheet and all capabilities of it are present and disposable.

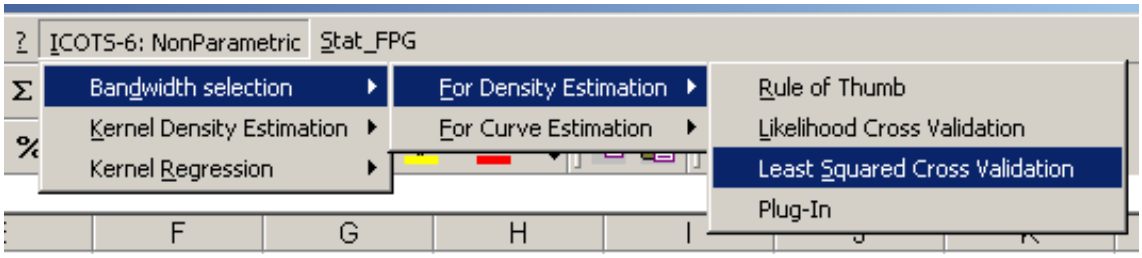

Figure 2. Pop up Menu Bar 1.

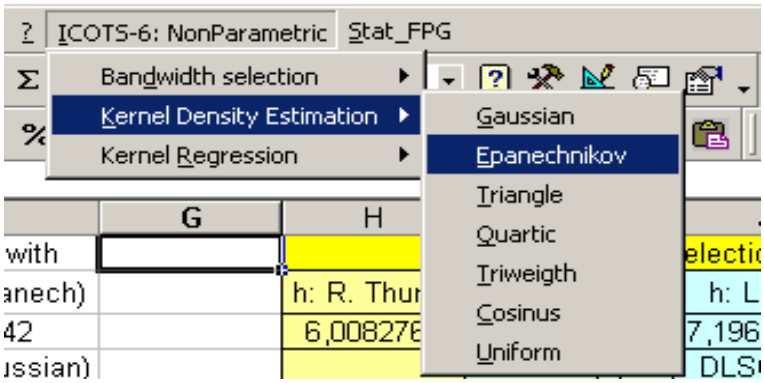

Figure 3. Pop Up Menu Bar 2.

According to the option selected into *ICOTS-6: Nonparametric*, one of the two UserForms is shown in Figures 4 and 5. In both figures, the range of the worksheet containing the data must be specified into the data box. In Figure 4, the range of grid on which we want to estimate the density or regression curve must be specified into the second box, while both the bandwidth *h* and the place of the sheet for allocating the results are placed into the third and fourth boxes respectively. In Figure 5, both the kernel and the allocation of results into the worksheet must be specified into the second and third boxes respectively. Previously, the data must be loaded into the spreadsheet and the desired grid must be specified into a place on the worksheet. The rest of the work is done by the package routine. The package has total flexibility, that is, at any time, the outputs can be adapted to any didactic contingency by the expert teacher, for example, if partial results are required, these can be shown into the worksheet. Besides this, the usual Dispersion Charts of Excel can be used to visualize and analyze results. I would like to encourage statistics teachers with Visual Basic knowledge to prepare similar programs for their particular needs.

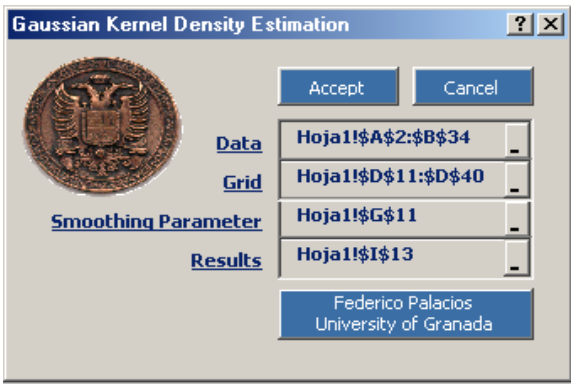

Figure 4. UserForm Worksheet 1.

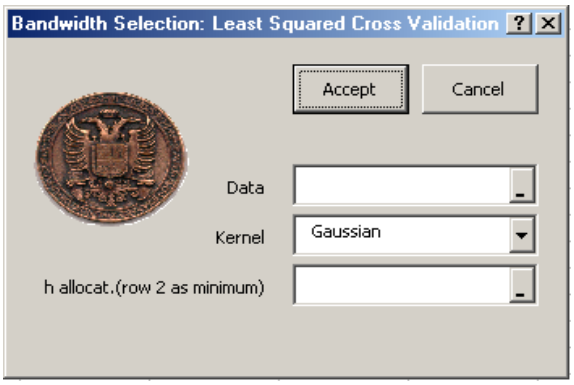

Figure 5. UserForm Worksheet 2.

REFERENCES.

Beauchamp Y., & Youssef, Y.A. (1998). An effective approach to teach design of experiments (DOE) using calculation-and-analysis worksheets and computerized spreadsheets. *Computers in Engineering. 35*(3/4), 643-646.

Bell, P.C. (2000). Teaching business statistics with Microsoft excel. *INFORMS Transactions on Education 1*(1), 18-26. Available at http://ite.informs.org/Vol1No1/bell/bell.html .

Cao, R., Cuevas A., & González Manteiga, W.G. (1994). A comparative study of several smoothing methods in density estimation. *Computational Statistics & Data Analysis*, *17,* 153-176.

Eubank, R.L. (1988). *Spline smoothing and nonparametric regression*. New York: Marcel Dekker.

Evans, J.R. (2000). Spreadsheets as a tool for teaching simulation. *INFORMS Transactions on Education 1*(1), 27-37. http://ite.informs.org/Vol1No1/evans/evans.html .

Farmen, M., & Marron, J.S. (1999). An assessment of finite sample performance of adaptative methods in density estimation. *Computational Statistics & Data Analysis, 30,* 143-168.

Hall, A.G. (1995). A Workshop approach using spreadsheets for the teaching of statistics and probability. *Computers in Engineering 25*(1/2), 5-12.

Härdle, W. (1991). *Smoothing techniques with implementation in S*. New York: Springer-Verlag.

Horgan, G.W. (1999). Use of spreadsheets for demonstrating experimental power and variability. *J. Statistics Education,* 7-1. http://www.amstat.org/publications/jse/secure/v7n1/horgan.cfm

Hunt, N. (1994). Teaching statistical concepts using spreadsheets. In *Proceedings of the 4th Conference on Teaching Statistic,* 2, 43. www.mis.coventry.ac.uk/~nhunt/ASLU.htm.

Hunt, N. (1996). *Teaching Statistics with Excel 5.0.* http://www.stats.gla.ac.uk/cti/activities/reviews/96_05/excel.html.

Gatti, G.G., & Harwell, M. (1998). Advantages of computer programs over powert charts for the estimation of power. J. *Stats Educ.*, 6-3. www.amstat.org/publications/jse/v6n3/gatti.html

Nash, J.C., & Quon, T.K. (1996). Issues in Teaching Statistical Thinking with Spreadsheets. *Journal of Statistics Education*, 4-1. http://www.amstat.org/publications/jse/v4n1/nash.html.

Jones, M.C., Marron, J.S., & Sheather, S.J. (1996). A brief survey of bandwidth selection for density estimation. *Journal of American Statistical Association, 87*, 227-233.

Loader, C.R. (1999). Bandwidth selection: Classical or Plug-In? *The Annals of Statistics, 27*(2), 415-438.

Park, B.U., & Marron, J.S. (1990). Comparison of data-driven bandwidth selectors. *Journal of American Statistical Association, 85,* 66-72.

Park, B.U., & Turlach, B.A. (1992). Practical performance of several data driven bandwidth selectors. Computational Statistics. *Computational Statistics, 7*, 251-270

Sheather, S.J. (1992). The performance of six popular bandwidth selection methods on some real datasets. *Computational Statistic, 7,* 225-250.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis.* New York: Chapman & Hall.