THE USE OF STATISTICAL SOFTWARE TO TEACH NONPARAMETRIC CURVE ESTIMATION: FROM EXCEL TO *R*

Ricardo Cao and Salvador Naya
Research Group MODES, Department of Mathematics, University of A Coruña, Spain
rcao@udc.es

The advantages of using R and Excel for teaching nonparametric curve estimation are presented in this paper. The use of these two tools for teaching nonparametric curve estimation is illustrated by means of several well-known data sets. Computation of histogram and kernel density estimators as well as kernel and local polynomial regression estimators is presented using Excel and R. Interactive changes in the sample and the smoothing parameter are illustrated using both tools. R incorporates sophisticated routines for crucial issues in nonparametric curve estimation, as smoothing parameter selection. The paper concludes summarizing the relative merits of these two tools for teaching nonparametric curve estimation and presenting RExcel, a free add-in for Excel that can be downloaded from the R distribution network.

INTRODUCTION

There has been an enormous expansion, over the past few years, on the use of computer and communication technologies for teaching statistics at different levels.

Microsoft Excel is the most popular spreadsheet program that is used to store information in columns and rows, which can then be organized and/or processed. Many authors consider Microsoft Excel as an excellent tool for statistical education (see, for example, Giles, 2002).

On the other hand, the use of free software is also one of the most interesting available tools for teaching statistics. Universal access to the Internet enables easy installation of free software. The use of large databases and electronic books concerning this software is also a common practice via the Internet. Discussion forums about the use of this software are common tools for getting trained on them. *R* is a free and open source environment. It is one of the most powerful and the fastest-growing statistical programs. It is very popular among researchers in statistics.

In between the spreadsheet Excel and the statistical package *R*, there has been recently emerged RExcel, a free add-in for Excel. It integrates the entire set of *R*'s statistical and graphical methods into Excel.

Only some of the statistical packages include routines for nonparametric curve estimation. On the other hand there is not much software to teach these techniques. Among the available material we mention the paper by Marron, Ruppert, Smith & Conley (2000) and the videos at Steve Marron's webpage http://www.unc.edu/~marron/Movies/locpoly_movies.html.

The statistical package *R* includes the library KernSmooth (Copyright by M. P. Wand 1997) to perform nonparametric kernel density and regression estimation. However, Microsoft Excel incorporates very few tools for nonparametric curve estimation: not much more than histograms and empirical distribution functions.

The pros and the cons of using Excel and *R* to teach nonparametric curve estimation are analyzed in this work. The main aim of this paper is to identify the possibilities of these two software tools and to give a critical review about their usefulness for teaching nonparametric curve estimation based on practical examples. The rest of the paper proceeds as follows. Section 2 presents the main features of Excel and its usefulness as a teaching tool. Section 3 is devoted to the statistical package *R*. The description of RExcel is included in Section 4. Section 5 includes two sample activities: a proposal for teaching nonparametric density estimation using Excel and *R* and another one for nonparametric regression. Finally, the conclusion of the paper is included in Section 6.

EXCEL AS A TEACHING TOOL

The spreadsheet is a powerful tool for teaching (see, for instance, Lewis (2006)). It is a helpful teaching resource for modelling, understanding and solving statistical problems. Unfortunately there are still plenty of teachers and students only familiar with its basic functions

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia.* Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

as tabular information and formula calculations. Other functionalities, not always commonly used, as dynamic tables, graphical outputs and simulations are very useful for students to make a bridge between intuitive ideas and formal concepts.

An important advantage of the spreadsheet Excel is that it has become a standard software in the teaching, professional and family environments. Excel has a friendly interface and it is easy to use.

Among the recent books and papers devoted to teaching statistics with Excel we mention the works by Lind, Marchal and Wathen (2004), Dretzke and Heilman (1998) and Teixeira, Rosa and Calapez (2009). However many statisticians consider Excel statistical tools too limited and with accuracy problems (see, for instance, McCullough & Wilson, 2005).

THE STATISTICAL PACKAGE *R*

R is a language and an environment for statistical computing and graphics (see R Development Core Team, 2007). It is a GNU project which is similar to the *S* language and environment which was developed at Bell Laboratories (now Lucent Technologies) by John Chambers and colleagues (see Ihaka & Gentleman, 1996). *R* can be considered as a different implementation of *S* (see Venables & Ripley, 2002, for a nice book about teaching statistics using *S*). *R* is a statistical system based on commands with its own programming syntax. It is possible to get quite far using *R* interactively, executing simple expressions from the command line. Some students may never need to go beyond that level, others will want to write their own functions either in an ad hoc fashion to systematize repetitive work or with the perspective of writing add-on packages for new functionality.

The books by Dalgaard (2008), Ugarte, Militino and Arnholt (2008) and Verzani (2005) are nice resources for teaching statistics using *R*. However the use of *R* is still limited for teaching statistics at a basic level due to fact that students need to learn the programming language.

Interfaces as *R* Commander (see Fox, 2005) have made *R* more friendly for its general use. *R* Commander uses a simple and familiar menu/dialog-box interface.

REXCEL

RExcel is a free add-in for Excel that can be downloaded from the *R* distribution network. Data can be transferred between R and Excel "the Excel way" by selecting worksheet ranges and using Excel menus. *R*'s basic statistical functions and selected advanced methods are available from an Excel menu. Results of the computations and statistical graphics can be returned back into Excel worksheet ranges. RExcel allows the use of Excel scroll bars and check boxes to create and animate *R* graphics as an interactive analysis tool. All *R* functions can be executed directly from within Excel worksheet formulas (including automatic recalculation when data change).

RExcel allows students to focus on statistical methods and concepts and to minimize the distraction of learning a new programming language. The book by Heiberger and Neuwirth (2009) is a very nice tool for teaching statistics with RExcel.

SAMPLE ACTIVITIES

Below we present descriptions of two sample activities in order to provide some examples with real data. We have chosen two well-known data sets to illustrate these activities. The first is the Rainfall Data (see McNeil, 1977). The second data set is the popular Motorcycle Accident Data presented in Silverman (1985).

Sample activity 1: Nonparametric density estimation

The first data set is concerned with the average amount of precipitation (rainfall) in inches for each of 70 United States (and Puerto Rico) cities. We use these data to illustrate nonparametric density estimation with Excel and *R*.

Students are asked to import the data in Excel, store them in a column, compute the cumulative and noncumulative histogram (using the function frequency) and plot them. A nonparametric kernel density estimator (see Silverman, 1986) can also be computed in Excel. The students are asked to fix a value x, and to compute the kernel estimator at x. To do this, it is convenient to define subsequent columns with the computations needed for every summand in the

definition of the estimator. This includes a final column with the evaluation of the kernel function. The sum of the last column gives the final value of the kernel density estimator at *x*. A dynamic table can be constructed to systematically repeat the same calculations on a grid of possible values for *x*. These values will be used to plot the estimated density.

Students are also asked to store the input value of the smoothing parameter in a prescribed Excel cell. Changing the value of this cell will produce new kernel density estimations with different bandwidths. This will be very useful for the students to learn the practical effect of the smoothing parameter choice. The effect of changing the kernel function can be also explored by changing the function used for the evaluation of the last column in the Excel calculations.

To perform similar actions in *R*, the students are asked to download the library KernSmooth. The histogram can be computed using the function hist. The function density can be used in *R* to compute the Parzen-Rosenblatt nonparametric density estimation, which can be easily plotted. The kernel function and the smoothing parameter are arguments in the function density. As a consequence, students can easily see the impact of changing the kernel and the bandwidth. This is very useful for understanding the practical meaning of undersmoothing and oversmoothing.

Students are asked to jointly plot the histogram and the kernel density estimation (see Figure 1.a). This is an easy way of understanding the smoothing effect of the kernel estimator. A similar technique is also useful to see the influence of the kernel function in the final estimation (see Figure 1.b).

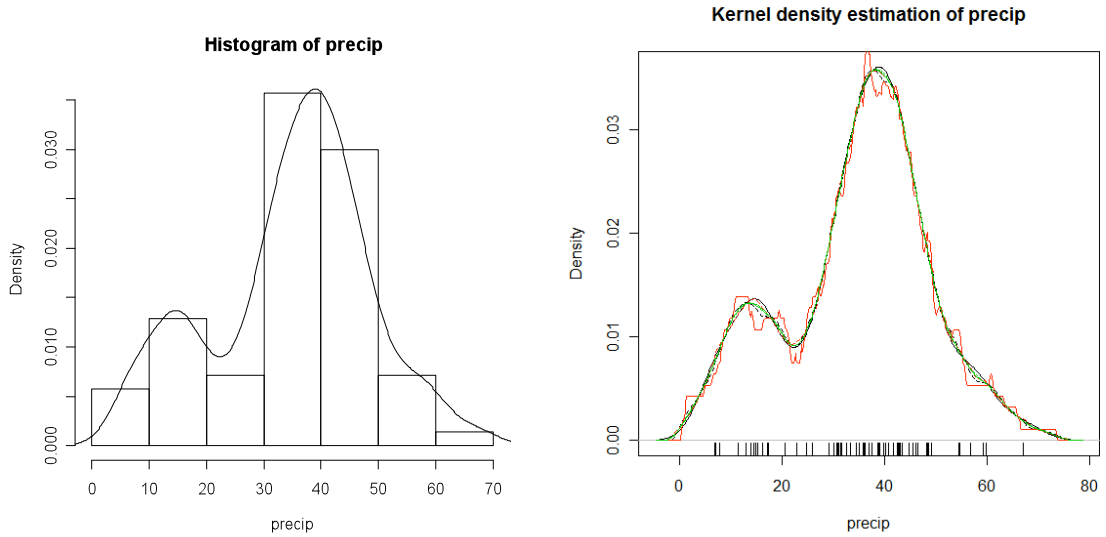

Figure 1. a) Histogram and kernel density estimation for the Rainfall data (left panel);
b) Kernel density estimation for the Rainfall data using different kernels (right panel)

The effect of changing the bandwidth is presented to the students by asking them to compute the kernel density estimation using three different values for the smoothing parameter: 1, 5 and 10 (see Figure 2.a). Students are also asked to use some data driven smoothing parameter selector, like the one proposed by Sheather and Jones (1991). This is easily done by introducing the value “sj” in the bandwidth argument of the *R* function density. Figure 2.b plots the final estimation using the Sheather & Jones bandwidth selector. Five different bandwidth selectors are available for this routine. Such sophisticated functions for bandwidth selection are not incorporated in Excel.

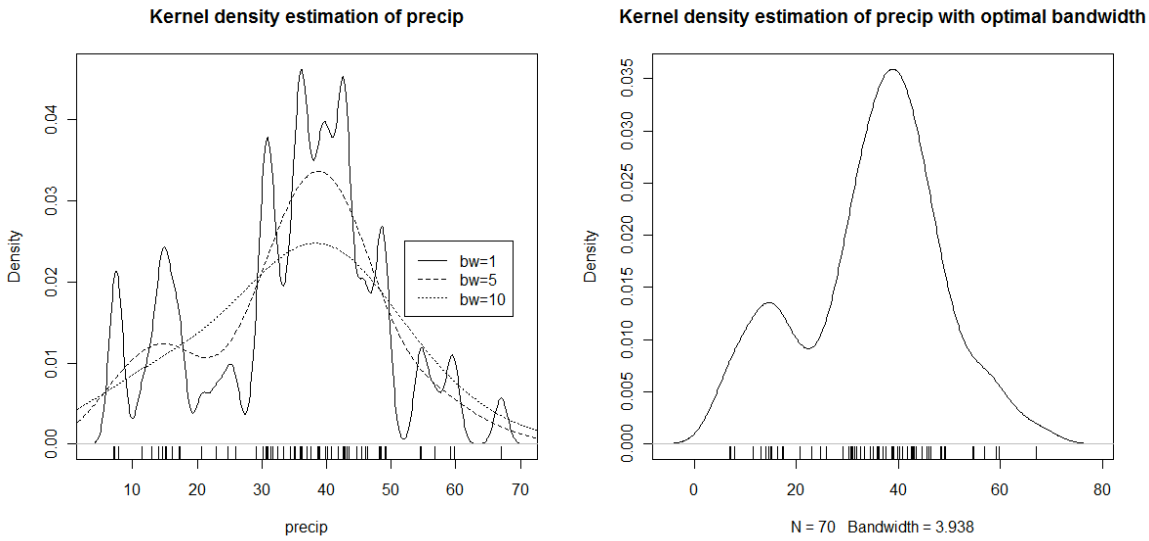

Figure 2. a) Kernel density estimation for the Rainfall data with different bandwidths: 1, 5 and 10 (left panel); b) Kernel density estimation for the Rainfall data with the Sheather & Jones bandwidth (right panel).

Sample activity 2: Nonparametric regression estimation

The Motorcycle Accident data set is used for this activity. The data contain a series of measurements of head acceleration in a simulated motorcycle accident, used to test crash helmets.

Students are asked to import the data in Excel and to select the option chart in the menu to plot them. To perform a parametric fit students will use the option add trendline, in the previous plot, to show a polynomial fit of the desired degree: 3 and 5 in this case. Nonparametric regression estimators, as the regressogram or the Nadaraya-Watson kernel estimator (see Nadaraya, 1964; Watson, 1964) are not directly incorporated in Excel. Students can compute these estimators following similar steps as done for the kernel density estimator presented in activity 1.

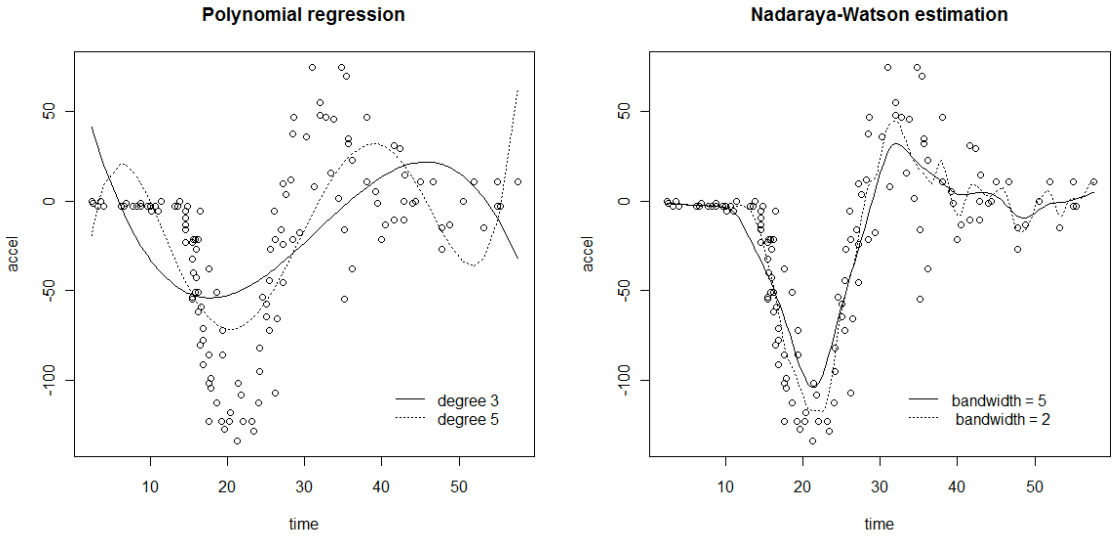

Figure 3. a) Polynomial estimation of degrees 3 and 5 for the Motorcycle data (left panel); b) Nadaraya-Watson kernel estimation for the Motorcycle data using two bandwidths: 2 and 5 (right panel).

The package *R* will be used by the students to produce polynomial fits (lm(y~poly(x,3)) function), with degrees 3 and 5, and a Nadaraya-Watson estimation (npreg function) for these data. Students are also asked to use the functions sm.spline (in the pspline package) to compute

the spline estimation and locpoly (in the KernSmooth package) for the local polynomial estimation with degree 1 (local linear estimation). The function dpill (in the KernSmooth package) is used to select the bandwidth of a local linear Gaussian kernel regression estimate using the direct plug-in methodology, as described by Ruppert, Sheather and Wand (1995). All these *R* functions cannot be found in Excel.

Students are asked to plot the polynomial estimation (see Figure 3.a) and the Nadaraya-Watson estimation (see Figure 3.b) for the Motorcycle data. This helps students to understand the flexibility of the nonparametric fits. Spline estimation (see Figure 4.a) and local polynomial estimation (see Figure 4.b) are also plotted by the students to compare the results of different nonparametric procedures.

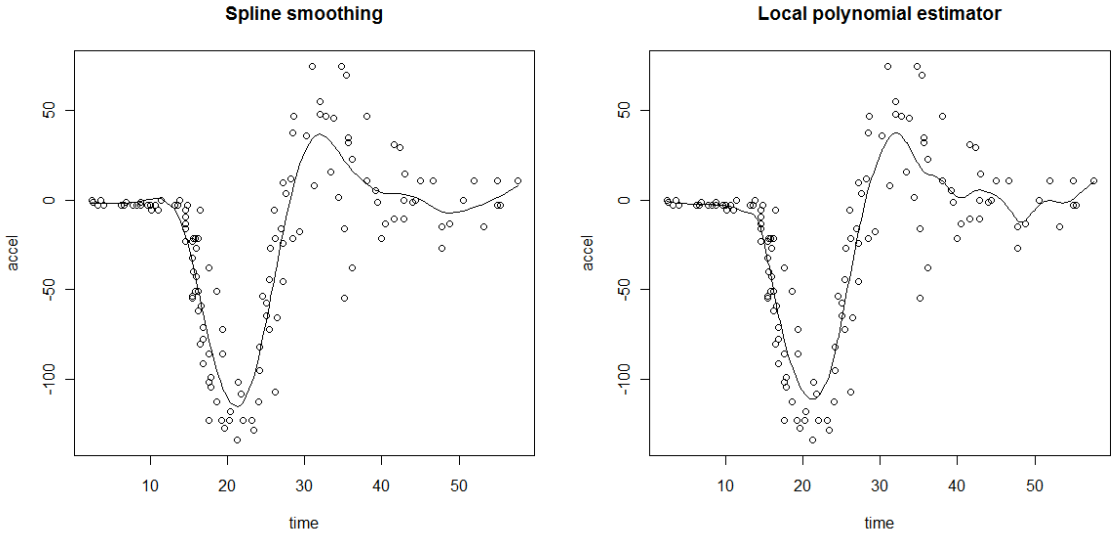

Figure 4. a) Spline estimation for the Motorcycle data (left panel);
b) Local polynomial estimation of degree 1 for the Motorcycle data using the direct plug-in bandwidth selector by Ruppert, Sheather & Wand (1995) (right panel).

CONCLUSION

While Microsoft Excel is a standard software that has become very popular among teachers in statistics, it has serious limitations for teaching specific fields within statistics, as nonparametric curve estimation. On the other hand the statistical environment *R* is an open source software, broadly used among researchers in statistics. However it is based on commands and it is oriented to users with some programming skills.

Nonparametric curve estimation includes rather standard techniques, as histograms or kernel density estimation, as well as some other sophisticated procedures, as those for bandwidth selection or local polynomial fitting. The latter are not available in Excel and time consuming to be implemented in Excel. The use of *R*, via a friendly interface, as *R* Commander, is a nice solution for teaching nonparametric curve estimation. Another recent possibility is RExcel, an Excel add-in that permits to use *R* within the Excel spreadsheet.

We propose to teach nonparametric curve estimation by developing materials to support an active learning pedagogical style. Some of the key features of these materials are illustrated in common elements of the sample activities presented above, including:

- Students conduct investigations of statistical concepts in nonparametric curve estimation.
- Estimators are introduced in the context of statistical ideas, applied to real data.
- Limitation of parametric techniques is exhibited and flexibility of nonparametric procedures is presented.
- Computer software is used as a tool for such techniques to assist with graphical displays and investigating effects of parameter changes, as for bandwidth influence.

We believe that this material will provide a more balanced way to introduce statistical concepts and methods in nonparametric curve estimation, serving as a bridge between intuitive ideas and more elaborated mathematical concepts.

ACKNOWLEDGMENTS

This work was partially supported by MICINN Grant MTM2008-00166 for both authors and XUGA Grant 07SIN012105PR for the first author.

REFERENCES

Dalgaard, P. (2008). *Introductory Statistics with R*. New York: Springer.

Dretzke, B.J. & Heilman, K.A. (1998). *Statistics with Microsoft Excel*. New Jersey: Prentice Hall.

Fox, J. (2005). The R Commander: A Basic-Statistics Graphical User Interface to R. *Journal of Statistical Software*, *14*.

Giles, O. (2002). Using excel to teach statistics in New Zealand secondary schools. ICOTS 2002.

Heiberger, R. M., & Neuwirth, E. (2009). *R Through Excel: A Spreadsheet Interface for Statistics, Data Analysis, and Graphics*. Series: Use R. New York: Springer.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5*, 299-314.

Lewis, P. (2006). *Spreadsheet Magic*. Washington: International Society for Technology in Education.

Lind, D. A., Marchal, W. G., & Wathen, S. A. (2004). *Basic Statistics Using Excel For Office Xp*. McGraw-Hill.

Marron, J. S., Ruppert, D., Smith, E. K., & Conley, G. (2000). Motion picture analysis of smoothing. *North Carolina Institute of Statistics, Mimeo Series* #2367.

McCullough, B. D., & Wilson, B. (2005). On the accuracy of statistical procedures in Microsoft Excel 2003. *Computational Statistics & Data Analysis*, *49*, 1244–1252.

McNeil, D. R. (1977). *Interactive Data Analysis*. New York: Wiley.

Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability and its Applications*, *9*, 141–142.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Online: www.R-project.org.

Ruppert, D., Sheather, S. J., & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, *90*, 1257–1270.

Sheather, S. J., & Jones M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society Series B*, *53*, 683–690.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric curve fitting. *Journal of the Royal Statistical Society Series B*, *47*, 1–52.

Silverman, B. W. (1986). *Density Estimation*. London: Chapman & Hall.

Teixeira, A., Rosa, A., & Calapez, T. (2009). Statistical power analysis with microsoft excel: normal tests for one or two means as a prelude to using non-central distributions to calculate power. *Journal of Statistical Education*, *17*.

Ugarte, M. D., Militino, A. F., & Arnholt, A. T. (2008). *Probability and Statistics with R*. Boca Raton: Chapman & Hall/CRC.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer.

Verzani, J. (2005). *Using R for Introductory Statistics*. Boca Raton: Chapman & Hall/CRC.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya*, *26*, 359–372.