The unexploited mines of academic and official statistics

Gilbert SAPORTA
Conservatoire National des Arts et Métiers, Chaire de Statistique Appliquée
292 rue Saint Martin, 75141 Paris Cedex 03, France
e-mail: saporta@cnam.fr

Abstract. There is often a large difference between methods proposed by academic researchers and those used in official statistics (numerical and non-numerical multivariate methods, Bayesian techniques, non-parametrics, etc.). Conversely, there is also some lack of interest from academic statisticians for official statistics. We will analyse the difference of scope between official and academic statisticians and try to propose ways to fill the gap.

1. Introduction

Official and academic (i.e. mathematical) statisticians are often evolving in two separate worlds and have few opportunities to meet. Despite the fact that official statisticians have been trained by academic statisticians, in many countries their paths and careers are separate and as a result there is little communication between them.
This situation is unfortunate since both kinds of statisticians would benefit from cooperation, but often they do not know that! On one hand official statistics may provide stimulating research areas and on the other hand mathematical statisticians have developed many methods which could be advantageously used by official statisticians.

2. Academic and official statistics: some misunderstandings

Academic and official statistics have a long and distinct history: official statistics was born several thousands of years ago with the objective of counting with accuracy the resources (human and material) of states. Its objective of quantitative assessment of the 'state of the society and the economy' [1] is still valid but the users of official statistics are not only governments but have been enlarged to all economic agents.
Mathematical statistics began more recently in the 17th century with Pascal, De Moivre and Bernoulli and were associated with probability theory. Up to the end of the 19th century there were very few connections between official and mathematical statistics and the idea of using probability theory (i.e. samples instead of census) to get information from a population was considered heretical until the ISI session of 1925 [2] after years of

debates and controversies. The problem then became how to draw a sample instead of why.
This controversy is a very good example of the progress that can be obtained by using mathematical methodology, and of the misunderstanding between these two worlds, and also of the prominent role of an organisation gathering governmental and academic statisticians.
However, a kind of opposition between 'blue-collar statisticians' (government statisticians) and 'white-collar statisticians' (methodologists) is still vivid [3].
Mathematical statistics, or theoretical statistics, is often far from the real world: excessive or undue formalisation gives too often to this discipline a rather negative feeling to users, not only in economics but also in industry. They consider many results as useless refinements.
Academic statisticians are not always interested in official statistics for several reasons: they see in official statistics only the activity of collecting, controlling and editing data, a professional occupation which deserves respect, but which needs only know-how and no formalisation. As pointed in [3] 'those who work for governments behave very often as if no error or no uncertainty existed'. Furthermore for some academics, official statistics may suffer from political pressures.

3. Official statistics as a mine of problems

Actually, there are many problems in official statistics that could stimulate the interest of academic statisticians: algorithms for ensuring confidentiality, small area estimation, visualisation techniques etc.
The problem of accuracy of data is, of course, a central one and due to the complexity of the processes of sampling and estimation, requires mathematical and/or computational skills. See again [3] who writes:

'The first and the most obvious is to develop means of measuring error where no such measurements exist and to ensure that new measurement tools are designed to a standard which in turn is made widely known to the user constituencies. The rigour that white collar statisticians bring to their task can at first appear to be misplaced particularly by government statisticians used as they are to the compromises that their applied discipline has asked them to adopt. But in the end the absence of rigour is what has led criticisms to be levied at the official numbers and in many cases has left official statisticians unable to provide an acceptable reply'.

4. Existing tools which are underused in official statistics

Among the enormous production of academic research, let us point out some methods that could be used by and to the benefit of official statisticians and which are not so widely used at present, according to my knowledge. I will classify these methods according to the field of applications.

4.1. Sampling

The estimation of the size of a population is of crucial interest in industrial statistics, especially in countries where official information is missing and where the informal sector is important. Biometricians used for many years capture-recapture methods for estimating the size of animal populations. This method may be transposed in official statistics without difficulty, see [4], and could provide confidence intervals for many ratios involving the population size as denominator.

4.2. Exploratory analysis

Descriptive statistics are the first outputs published: graphics are generally limited to a few bar charts and pie charts or histograms, and Lorenz curves. The use of modern graphics such as box-plots, especially for comparing subpopulations, and density estimation instead of histograms should be promoted.
Multidimensional data analysis (principal components, correspondence analysis) is now well established: mappings using principal axes are communications tools widely used in industry and market research. They could also be advantageously used for regional comparisons in official statistics.
Methods for analysing textual data, see [5], are able to deal with open-ended questions by using stylometrics and multivariate techniques. They could be used for opinion surveys, leisure activity surveys etc.

4.3. Modelling

Econometricians are using generally sophisticated models based on strong probabilistic assumptions and with properties that are often valid only asymptotically. There could be advantages to using 'soft modelling' based on weaker assumptions such as the PLS (partial least squares) approach initiated by H. Wold [6] which is successfully used in chemometrics for ill-conditioned models. PLS techniques are competitors to maximum likelihood [7].
Classical models (and PLS, too) are explicit models in the sense that they provide equations where the main problem is to estimate parameters. Non-parametric models have been developed for many years but they do not seem to have been really applied to econometrics despite the efforts of their promoters (see, for instance, the web site *http://www.xplore-stat.de/)*. One drawback of non-explicit models was that until recently they were not conceived for prediction: now it is possible to use them not only for interpolation but also for extrapolation.
Computer intensive techniques such as neural networks for prediction are currently used in many fields of application: why not try them for economic forecasting? Of course they are 'black boxes' and do not provide deep understanding of phenomena, but for short-term prediction it seems that the most important thing is to get good predictions more than having the most comprehensible model.
In official statistics *a priori* information is plentiful and analysts use it implicitly. Why not do it explicitly and apply Bayesian techniques? For a long time Bayesian methodology was first a controversial topic and also very difficult to apply due to the difficulty of

programming posterior distributions but now the status of Bayesian analysis is well established and efficient methods for Bayesian computing are available [8].

4.4. Data mining

As D. Hand [9] writes '*Data mining is the process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest for the data base owner*'. This new discipline is actually a merge of techniques coming from databases and statistics where some tools like cluster analyisis, decision trees and Bayesian networks are extensively used. The main field of application is, for the moment, customer and transactions databases of large commercial companies. As far as NSIs manage large databases on population, trade, agriculture and companies, they certainly would take a great profit of exploiting their data mines.
In the same context, there are recent advances on data fusion (merging files from different sources with large-scale estimation when questions are missing for some files) which could be exploited in official statistics [10].

5. Conclusion: some proposals to fill the gap

First of all, it is necessary that official and academic statisticians meet in common organisations. At the international level, the role of the International Statistical Institute (ISI) is very important, but at the national level both groups should participate in the same national statistical society. Common meetings where official statisticians expose their problems to academic ones are very useful. At the European level, Eurostat organises several scientific meetings each year: they are usually intended for official statisticians and academic researchers involved in research programs: the diffusion could be enlarged to the whole community of statisticians.
As we have pointed out earlier, academics are not aware of the problems of official statistics, especially mathematical statisticians: publishing books about methodological issues of official statistics could be useful.
Developing institutional links between National Statistical Institutes (NSIs) and universities should be encouraged: NSIs could give doctoral scholarships with a common supervision and help to introduce topics about official statistics in academic curriculums. One could also develop research contracts between NSIs and universities and foster exchanges between people: for instance, CREST (the research centre of the French NSI) welcomes university professors for temporary positions (two or three years) but a reciprocal arrangement does not exist.

References

[1] Cheung, P., 'Developments in official statistics and challenges for statistical education', *ICOTS 5*, Singapore, 1998.

[2] Droesbeke, J. J. and Tassi, P., 'Histoire de la Statistique', *Que Sais-je*, Presses Universitaires de France, Paris, 1990.

[3] J.Ryten, 'Blue and white collar statisticians, a gap revisited', *Conference of European Statisticians, Seminar on Official Statistics - Past and Future*, Lisbon, Portugal, 25-27 September 1996.

[4] Giommi et al., 'On the use of capture mark recapture methodology in estimating the size of an open population of firms', contributed paper, Vol. I, *ISI Session*, Peking, 1995.

[5] Lebart, L., Salem, A. & Berry, L., *Exploring textual data*, Kluwer Academic, Dodrecht, Netherlands, 1998.

[6] Jöreskog, K. G. & Wold, H., *Contibutions to Economic Analysis. Systems under indirect observation: causality, structure, prediction*, North Holland, Amsterdam, 1982.

[7] Tenenhaus, M., *La régression PLS : Théorie et Pratique*, Editions Technip, Paris, 1998.

[8] Robert, C., *The Bayesian Choice*, Springer, New York, 1994.

[9] Hand, D., 'Data Mining: Statistics and More?', *The American Statistician*, Vol. 52, No 2, pp. 112-118, 1998.

[10] Saporta, G. & Co, V., 'Data Fusion: A New Method Based on Homogeneity Analysis', *8th International Symposium Applied Stochastic Models and Data Analysis*, contributed papers, 1997, pp. 395-399.