The interplay of relevance and generalization in Biostatistics

J.-D. Lebreton [a]

[a] CEFE/CNRS,
route de Mende,
Montpellier, France
jean-dominique.lebreton@cefe.cnrs.fr

Abstract: In the development of mathematical methods, two contrasted, nearly contradictory logics are in action. This is particularly the case in Biomathematics, including Biostatistics and Statistical Ecology, which will be the main focus of my reflections and illustrations.

The logic of relevance stems from a full acceptation of biological questions, and then attempts at developing tools closely fitting these questions. After an initial tool is proposed, such as the probit model to analyse dose-response relationships, one generally see a proliferation of particularizations and variants, published one at a time, and often named from their author or by some exotic name. Many fields can serve as illustrations of this proliferation process: predator-prey models, capture-recapture methods, descriptive multivariate analysis ("data analysis"), etc… After an initial success, this proliferation of methods (and software) is often a source of confusion for users, with little help from a poor nomenclature. Another clear risk of the logic of relevance is of developing "ad hoc", statistically non optimal approaches to any particular question, that may become the dominant practices for years within a particular scientific community.

The logic of generalization comes from pure mathematics, and is based on the idea that, by generalizing an existing mathematical object, you will unavoidably visit new, unexplored, territories, discover unexpected links, and make valuable encounters. The use of the duality diagram as a prospective tool in descriptive multivariate analysis and of generalized linear models as a common frame for a variety of discrete data models are obvious examples of the logic of generalization. In face of the advantage of unifying existing approaches and opening new avenues of development, the clear risk is to use a sledgehammer to crack a nutshell, or, worse, to use a hammer to fit a screw. Using a fancy mixed logistic model for estimating survival from data on marked individuals, not accounting for incomplete detection which is the key feature of such data, would be an example of such a mismatch.

One can easily deduce from such premises that statistical Ecology, and biomathematics in general, could not have survived, developed, and be efficient and useful with only one of these two logics at work. I will show and illustrate how these two logics fit together in successive phases of development, each one needing an accumulation of material from the other one to be fully efficient. In a pluridisciplinary endeavour such as statistical ecology, the reflections should necessarily encompass the development of software and shared data bases. I will go on discussing the strategies of research and transfer of knowledge that can be thought of in such a framework.