Foreword

Gilbert Saporta

Foreword

It is a great pleasure for me to preface this imposing work which establishes, with *Analysis of Symbolic Data* (Bock and Diday, 2000) and *Symbolic Data Analysis* (Billard and Diday, 2006), a true bible as well as a practical handbook.

Since the pioneering work of Diday at the end of the 1980s, symbolic data analysis has spread beyond a restricted circle of researchers to attain a stature attested to by many publications and conferences. Projects have been supported by Eurostat (the statistical office of the European Union), and this recognition of symbolic data analysis as a tool for official statistics was also a crucial step forward.

Symbolic data analysis is part of the great movement of interaction between statistics and data processing. In the 1960s, under the impetus of J. Tukey and of J.P. Benzécri, exploratory data analysis was made possible by progress in computation and by the need to process the data which one stored. During this time, large data sets were tables of a few thousand units described by a few tens of variables. The goal was to have the data speak directly without using overly simplified models. With the development of relational databases and data warehouses, the problem changed dimension, and one might say that it gave birth on the one hand to data mining and on the other hand to symbolic data analysis. However, this convenient opposition or distinction is somewhat artificial.

Data mining and machine learning techniques look for patterns (exploratory or unsupervised) and models (supervised) by processing almost all the available data: the statistical unit remains unchanged and the concept of model takes on a very special meaning. A model is no longer a parsimonious representation of reality resulting from a physical, biological, or economic theory being put in the form of a simple relation between variables, but a forecasting algorithm (often a black box) whose quality is measured by its ability to generalize to new data (which must follow the same distribution). Statistical learning theory provides the theoretical framework for these methods, but the data remain traditional data, represented in the form of a rectangular table of variables and individuals where each cell is a value of a numeric variable or a category of a qualitative variable assumed to be measured without error.

Symbolic data analysis was often presented as a way to process data of another kind, taking variability into account: matrix cells do not necessarily contain a single value but an interval of values, a histogram, a distribution. This vision is exact but reductive, and this book shows quite well that symbolic data analysis corresponds to the results of database operations intended to obtain conceptual information (knowledge extraction). In this respect symbolic

data analysis can be considered as the statistical theory associated with relational databases. It is not surprising that symbolic data analysis found an important field of application in official statistics where it is essential to present data at a high level of aggregation rather than to reason on individual data. For that, a rigorous mathematical framework has been developed which is presented in a comprehensive way in the important introductory chapter of this book.

Once this framework was set, it was necessary to adapt the traditional methods to this new type of data, and Parts II and III show how to do this both in exploratory analysis (including very sophisticated methods such as generalized canonical analysis) and in supervised analysis where the problem is the interrelation and prediction of symbolic variables. The chapters dedicated to cluster analysis are of great importance. The methods and developments gathered together in this book are impressive and show well that symbolic data analysis has reached full maturity.

In an earlier paragraph I spoke of the artificial opposition of data mining and symbolic data analysis. One will find in this book symbolic generalizations of methods which are typical of data mining such as association rules, neural networks, Kohonen maps, and classification trees. The border between the two fields is thus fairly fluid.

What is the use of sound statistical methods if users do not have application software at their disposal? It is one of the strengths of the team which contributed to this book that they have also developed the freely available software SODAS2. I strongly advise the reader to use SODAS2 at the same time as he reads this book. One can only congratulate the editors and the authors who have brought together in this work such an accumulation of knowledge in a homogeneous and accessible language. This book will mark a milestone in the history of data analysis.

Gilbert Saporta