Studies in Classification, Data Analysis, and Knowledge Organization

Managing Editors
H.-H. Bock, Aachen
W. Gaul, Karlsruhe
M. Vichi, Rome

Editorial Board
Ph. Arabie, Newark
D. Baier, Cottbus
F. Critchley, Milton Keynes
R. Decker, Bielefeld
E. Diday, Paris
M. Greenacre, Barcelona
C. Lauro, Naples
J. Meulman, Leiden
P. Monari, Bologna
S. Nishisato, Toronto
N. Ohsumi, Tokyo
O. Opitz, Augsburg
G. Ritter, Passau
M. Schader, Mannheim
C. Weihs, Dortmund
Titles in the Series

M. Schader (Ed.)
Analyzing and Modeling Data and Knowledge. 1992

O. Opitz, B. Lausen, and R. Klar (Eds.)
Information and Classification. 1993 (out of print)

H.-H. Bock, W. Lenski, and M.M. Richter (Eds.)
Information Systems and Data Analysis. 1994 (out of print)

E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy (Eds.)
New Approaches in Classification and Data Analysis. 1994 (out of print)

W. Gaul and D. Pfeifer (Eds.)
From Data to Knowledge. 1995

H.-H. Bock and W. Polasek (Eds.)
Data Analysis and Information Systems. 1996

E. Diday, Y. Lechevallier, and O. Opitz (Eds.)
Ordinal and Symbolic Data Analysis. 1996

R. Klar and O. Opitz (Eds.)
Classification and Knowledge Organization. 1997

C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock, and Y. Baba (Eds.)
Data Science, Classification, and Related Methods. 1998

I. Balderjahn, R. Mathar, and M. Schader (Eds.)
Classification, Data Analysis, and Data Highways. 1998

A. Rizzi, M. Vichi, and H.-H. Bock (Eds.)
Advances in Data Science and Classification. 1998

M. Vichi and O. Opitz (Eds.)
Classification and Data Analysis. 1999

W. Gaul and H. Locarek-Junge (Eds.)
Classification in the Information Age. 1999

H.-H. Bock and E. Diday (Eds.)
Analysis of Symbolic Data. 2000

H.A.L. Kiers, J.-P. Rasson, P.J.F. Groenen, and M. Schader (Eds.)
Data Analysis, Classification, and Related Methods. 2000

W. Gaul, O. Opitz, and M. Schader (Eds.)
Data Analysis. 2000

R. Decker and W. Gaul (Eds.)
Classification and Information Processing at the Turn of the Millenium. 2000

S. Borra, R. Rocci, M. Vichi, and M. Schader (Eds.)
Advances in Classification and Data Analysis. 2001

W. Gaul and G. Ritter (Eds.)
Classification, Automation, and New Media. 2002

K. Jayuga, A. Sokolowski, and H.-H. Bock (Eds.)
Classification, Clustering and Data Analysis. 2002

M. Schwaiger and O. Opitz (Eds.)
Exploratory Data Analysis in Empirical Research. 2003

M. Schader, W. Gaul, and M. Vichi (Eds.)
Between Data Science and Applied Data Analysis. 2003

H.-H. Bock, M. Chiodi, and A. Mineo (Eds.)
Advances in Multivariate Data Analysis. 2004

D. Banks, L. House, F.R. McMorris, P. Arabie, and W. Gaul (Eds.)
Classification, Clustering, and Data Mining Applications. 2004

D. Baier and K.-D. Wernecke (Eds.)
Innovations in Classification, Data Science, and Information Systems. 2005
Preface

This volume contains revised versions of selected papers presented during the biannual meeting of the Classification and Data Analysis Group of Società Italiana di Statistica, which was held in Bologna, September 22-24, 2003.

The scientific program of the conference included 80 contributed papers. Moreover it was possible to recruit six internationally renowned invited speakers for plenary talks on their current research works regarding the core topics of IFCS (the International Federation of Classification Societies) and Wolfgang Gaul and the colleagues of the GfKl organized a session. Thus, the conference provided a large number of scientists and experts from home and abroad with an attractive forum for discussions and mutual exchange of knowledge.

The talks in the different sessions focused on methodological developments in supervised and unsupervised classification and in data analysis, also providing relevant contributions in the context of applications. This suggested the presentation of the 43 selected papers in three parts as follows:

CLASSIFICATION AND CLUSTERING
Non parametric classification
Clustering and dissimilarities

MULTIVARIATE STATISTICS AND DATA ANALYSIS

APPLIED MULTIVARIATE STATISTICS
Environmental data
Microarray data
Behavioural and text data
Financial data

We wish to express our gratitude to the authors whose enthusiastic participation made the meeting possible. We are very grateful to the reviewers for the time spent in their professional reviewing work. We would also like to extend our thanks to the chairpersons and discussants of the sessions: their comments and suggestions proved very stimulating both for the authors and the audience.

We also thank the Italian Ministero dell’Istruzione, dell’Università e della Ricerca for partially supporting the pubblication of this volume.
VI  Preface

Special thanks are also due to the local organizing committee: G. Galimberti, L. Guidotti, A. Lubisco, S. Mignani, P. Monari, A. Montanari, M. Pillati, G. Soffritti, to the University of Bologna for financial support and hospitality, and to the many persons who, in an unofficial form, strongly contributed to the organization of the conference.

We would finally like to thank Dr. M. Bihn and her colleagues from Springer-Verlag, Heidelberg, for the excellent cooperation in publishing this volume.

Roma, Bologna
October 2004

Maurizio Vichi
Paola Monari
Stefania Mignani
Angela Montanari
## Contents

### Part I Classification and Clustering

#### Non Parametric Classification

- **Multi-Class Budget Exploratory Trees**  
  *Massimo Aria* .......................................................... 3

- **Methods to Compare Nonparametric Classifiers and to Select the Predictors**  
  *Simone Borra, Agostino Di Ciaccio* .................................. 11

- **Variable Selection in Cell Classification Problems: a Strategy Based on Independent Component Analysis**  
  *Daniela G. Calò, Giuliano Galimberti, Marilena Pillati, Cinzia Virolı* . 21

- **Simplifying Classification Trees Through Consensus Methods**  
  *Rossella Miglio, Gabriele Soffritti* .................................. 31

- **Selecting the Training Set in Classification Problems with Rare Events**  
  *Bruno Scarpa, Nicola Torelli* ........................................ 39

#### Clustering and Dissimilarities

- **A Classification and Discrimination Integrated Strategy Conducted on Symbolic Data for Missing Data Treatment in Questionnaire Survey**  
  *Maria Gabriella Grussia* ................................................ 47

- **A Collinearity Based Hierarchical Method to Identify Clusters of Variables**  
  *Annalisa Laghi, Gabriele Soffritti* .................................. 55

- **On the Dynamic Time Warping for Computing the Dissimilarity Between Curves**  
  *Isabella Morlini* ........................................................ 63

- **Metrics in Symbolic Data Analysis**  
  *Luciano Nieddu, Alfredo Rizzi* ....................................... 71
## Constrained Clusterwise Linear Regression
*Antonella Plaza* .................................................. 79

## Crossed Clustering Method on Symbolic Data Tables
*Rosanna Verde, Yves Lechevallier* ............................. 87

---

### Part II Multivariate Statistics and Data Analysis

#### Some Statistical Applications of Centrosymmetric Matrices
*Estela Bee Dagum, Laura Guidotti, Alessandra Luati* ..................... 97

#### Selection of Structural Equation Models with the PLS-VB Programme
*Giuseppe Boari, Gabriele Cantaluppi* ........................................ 105

#### Web Robot Detection - Preprocessing Web Logfiles for Robot Detection
*Christian Bomhardt, Wolfgang Gaul, Lars Schmidt-Thieme* ............... 113

#### Data Dependent Prior Modeling and Estimation in Contingency Tables: The Order-Restricted RC Model
*Giulio D’Epifanio* ..................................................... 125

#### PLS Typological Regression: Algorithmic, Classification and Validation Issues
*Vincenzo Esposito Vinzi, Carlo N. Lauro, Silvano Amato* .................. 133

#### Weighted Metric Multidimensional Scaling
*Michael Greenacre* .......................................................... 141

#### An Improved Majorization Algorithm for Robust Procrustes Analysis
*Patrick J.F. Groenen, Patrizia Giaquinto, Henk A.L. Kiers* ............... 151

#### Generalized Bi-Additive Modelling for Categorical Data
*Patrick J.F. Groenen, Alex J. Koning* ......................................... 159

#### Model Selection Procedures in Three-Mode Component Models
*Pieter M. Kroonenberg* ....................................................... 167

#### Principal Component Analysis for Non-Precise Data
*Carlo N. Lauro, Francesco Palumbo* ........................................... 173

#### A New Version of the Structural Dynamic Model with Unique Latent Scores
*Simona Caterina Minotti, Giorgio Vittadini* ................................... 185

#### Some Issues About the Use of Experts
*Paola Monari, Patrizia Agati* .................................................. 193
Nonparametric Methods in Survey Sampling  
*Giorgio E. Montanari, M. Giovanna Ranalli* .................................. 203

A Different Approach for the Analysis of Web Access Logs  
*Gabriella Schoier, Giuseppe Melfi* ............................................. 211

Blending Statistics and Graphics in Visual Data Mining  
*Antony Unwin* ................................................................. 217

Confidence Regions in Multivariate Calibration: A Proposal  
*Diego Zappa, Silvia Salini* .................................................. 225

**Part III Applied Multivariate Statistics**

**Environmental Data**

Nonparametric Analysis of Air Pollution Indices Compositions  
*Francesca Bruno, Daniela Cocchi, Meri Raggi* .......................... 235

VAR Models for Spatio-temporal Structures: an Application to Environmental Data  
*Aldo Lamberti, Alessia Naccarato* ....................................... 243

Analysis of Spatial Covariance Structure for Environmental Data  
*Aldo Lamberti, Eugenia Nissi* .............................................. 251

**Microarray Data**

Bayesian Clustering of Gene Expression Dynamics: An Application  
*Sara Lodi, Paola Sebastiani, Daniela Cocchi, Marco Ramoni* .......... 259

A Hierarchical Mixture Model for Gene Expression Data  
*Luisa Scaccia, Francesco Bartolucci* .................................... 267

**Behavioural and Text Data**

Sequence Analysis of BHPS Life Course Data  
*Arnstein Aassve, Francesco C. Billari, Raffaella Piccarreta* .......... 275

Robust Multivariate Methods for the Analysis of the University Performance  
*Matilde Bini* ......................................................................... 285

Some Insights into the Evolution of 1990s’ Standard Italian Using Text Mining Techniques and Automatic Categorization  
*Sergio Bolasco, Alessio Canzonetti* ....................................... 293

Evaluating Undergraduate Education Through Multiple Choice Tests  
*Franca Crippa* ..................................................................... 303
A New Approach in Business Travel Survey: Multivariate Techniques for Strata Design
*Andrea Guizzardi* .................................................. 311

Value-Orientations and Partnership
*Rossella Miglio, Lucia Pasquini, Alessandra Samoggia, Gabriele Soffritti* .......................................................... 317

An Item Response Theory Model for Student Ability Evaluation Using Computer-Automated Test Results
*Stefania Mignani, Silvia Cagnone, Giorgio Casadei, Antonella Carbonaro* .......................................................... 325

**Financial Data**

Functional Cluster Analysis of Financial Time Series
*Andrea Cerioli, Fabrizio Laurini, Aldo Corbellini* .................. 333

The Role of the Normal Distribution in Financial Markets
*Michele Costa, Giuseppe Cavaliere, Stefano Iezzi* .................. 343

Functional Principal Component Analysis of Financial Time Series
*Salvatore Ingrassia, G. Damiana Costanzo* ......................... 351

Elliptically Symmetric Distributions: A Review of Achieved Results and Open Issues
*Fiorella Romito* ....................................................... 359

**Author and Subject Index** .......................................... 367
Classification and Clustering

Non parametric classification  3

Clustering and dissimilarities  47