

Classification and Data Analysis 2009

Book of Short Papers
7° Meeting of the Classification and Data Analysis Group
of the Italian Statistical Society



Catania – September 9-11, 2009

Editors

Salvatore Ingrassia and Roberto Rocci

cleup

This book was realized with the financial support of:

Dipartimento di Economia e Metodi Quantitativi, Università di Catania

Dipartimento di Sociologia e Metodi delle Scienze Sociali, Università di Catania

Dipartimento di Statistica "G. Parenti", Università di Firenze

Facoltà di Economia, Università di Catania

Facoltà di Scienze Politiche, Università di Catania

Università degli Studi di Catania

Prima edizione: luglio 2009

ISBN 978-88-6129-406-6

© Copyright 2009 by CLEUP sc
"Coop. Libreria Editrice Università di Padova"
Via G. Belzoni, 118/3 – Padova (Tel. 049 650261)
www.cleup.it

Tutti i diritti di traduzione, riproduzione e adattamento, totale o parziale, con qualsiasi mezzo (comprese le copie fotostatiche e i microfilm) sono riservati.

Table of contents

Preface.....XIX

KEYNOTE PAPERS

A. Cerioli
A guided tour of multivariate outlier detection with and without high-breakdown estimators..... 3

J. C. Gower
Extended canonical analysis: theory and applications..... 7

D. B. Rubin
Recent issues in causal inference 11

INVITED PAPERS

Specialized Session 1
Multivariate Relations for General Data
Organizer: Carles M. Cuadras

C. M. Cuadras
Distance-based approach in multivariate association..... 7

K. Gibert, G. Rodríguez-Silva, A. García-Rudolph, J. M. Tormos
On the benefits of mixed AI and statistics methods for KDD in real-world domains. Clustering based on rules and extensions 21

J. Graffelman
A universal procedure for biplot calibration..... 25

Specialized Session 2
**Estimating Causal Effects from Observational Data
 using Large Databases and Multivariate Matching**
 Organizer: Fabrizia Mealli

B. Arpino, F. Mealli
*The specification of the propensity score in multilevel
 observational studies* 31

F. Camillo, I. D'Attoma
*A multivariate approach to assess balance of categorical covariates in
 observational studies* 35

S. M. Iacus, G. King, G. Porro
Coarsened exact matching 39

Specialized Session 3
New Developments in Complex Data Modeling for Evaluation
 Organizer: Giorgio Vittadini

R. Brüggemann, M. Fattore, J. Owsinski
*Using poset theory to compare fuzzy multidimensional poverty across
 regions* 45

S. C. Minotti
*Some notes on the use of cluster-weighted modeling in effectiveness
 studies*..... 49

F. Pennoni, F. Bartolucci
Impact evaluation of job training programs by a latent variable model 53

Specialized Session 4
**Classification of Measurement Errors: Theory, ISO Standards
 and Applications**
 Organizer: Laura Deldossi, Diego Zappa

L. Deldossi, D. Zappa
Measurement errors and uncertainty: a statistical perspective 59

D. Koltai, R. S. Kenett and D. Kristt
*Measurement uncertainty in quantitative chimerism monitoring
 after stem cell transplantation* 63

S. Salini, N. Solaro	
<i>Controlled calibration in presence of clustered measures</i>	67

Specialized Session 5
Visualization of Relationships by Multidimensional Scaling
Organizer: Akinori Okada

G. Bove	
<i>Analysis of skew-symmetry in proximity data</i>	73
M. Nakai	
<i>Social stratification and consumption patterns: cultural practices and lifestyles in Japan</i>	77
A. Okada	
<i>Centrality of asymmetric social network: singular value decomposition, conjoint measurement, and asymmetric multidimensional scaling</i>	81

Specialized Session 6
Computational and Theoretical Aspects in Mixture Models
Organizer: Donatella Vicari

M. Alfò, L. Nieddu, D. Vicari	
<i>Regression models for multivariate mixed responses</i>	87
D. G. Calò, C. Viroli	
<i>A finite mixture model for multilevel data</i>	91
C. Hennig	
<i>How to merge Gaussian mixture components</i>	95

Specialized Session 7
Recent Trends in Cluster Analysis
Organizers: Yves Lechevallier

Z. Assaghir, M. Kaytoue, N. Messai, A. Napoli	
<i>On the mining of numerical data with formal concept analysis and similarity</i>	101

M. Chavent, V. Kuentz, J. Saracco <i>A partitioning method for the clustering of qualitative variables</i>	105
A. Hardy <i>Validation of a clustering structure for symbolic data</i>	109

Specialized Session 8
Models for the Analysis of Volatility in Financial Markets
Organizer: Edoardo Otranto

G. Barone-Adesi <i>Volatility dynamics: escaping the straitjacket</i>	115
P. Coretto, M. La Rocca, G. Storti <i>Group structured volatility</i>	119
G. Giovannetti, M. Velucchi <i>A MEM analysis of African financial markets</i>	123

Specialized Session 9
Model selection in Advanced Classification and Regression Methods
Organizers: Roberta Siciliano, Claudio Conversano

C. Conversano, E. Dusseldorp <i>Model selection in classification trunk approach</i>	129
A. D'Ambrosio, W. J. Heiser <i>Decision trees for preference rankings</i>	133
J. S. Rao, J. Jiang, T. Nguyen <i>A brief overview of fence methods for mixed model selection</i>	137

Specialized Session 10
Operative Problems of Data Mining and Statistical Solutions
Organizer: Furio Camillo

H. Bozdogan, J. A. Howe, S. Katragadda, C. Liberati <i>Misspecification resistant model selection using information complexity with applications</i>	143
---	-----

G. Galimberti, M. Pillati, G. Soffritti <i>Some issues on robustness of regression trees</i>	147
---	-----

G. Menardi, N. Torelli <i>Some issues in building and assessing classification rules with extremely skewed datasets</i>	151
--	-----

Specialized Session 11
**Statistical Methods for the Assessment of
Scientific Research Communities**

Organizer: Giuseppe Giordano

C. Davino, F. Palumbo, D. Vistocco <i>Environmental factors and scientific production: a statistical analysis of MUR data</i>	157
--	-----

D. De Stefano, M. P. Vitale <i>Exploring the pattern of scientific collaboration networks</i>	161
--	-----

L. Leydesdorff, Y. Sun <i>The development of scientific co-authorship networks in national and international contexts</i>	165
--	-----

Specialized Session 12
Stable Likelihood Methods in Mixture Models

Organizer: Wilfried Seidel

C. Biernacki <i>Managing degeneracy in Gaussian mixtures</i>	171
---	-----

S. Ingrassia, R. Rocci <i>Constrained EM trajectories for mixtures of normal distributions</i>	175
---	-----

K. Sever <i>Dealing with spurious solutions in finite mixture models</i>	179
---	-----

Specialized Session 13
Statistical Analysis of Microarray Data
Organizer: Angelo M. Mineo

M. Alfó, F. Martella <i>Identifying partitions of genes and tissue samples in microarray data</i>	185
L. Augugliaro, E. C. Wit <i>Generalizing LARS algorithm using differential geometry</i>	189
E. C. Wit, L. Augugliaro <i>Mixed modelling ideas for microarray data</i>	193

Specialized Session 14
Robust Methods for the Analysis of Complex Data
Organizer: Marco Riani

T. DiCiccio, A. C. Monti <i>Flexible models and robustness</i>	199
L. Grossi, F. Laurini <i>Robust portfolio asset allocation</i>	203
A. Marazzi <i>Robust regression with asymmetric errors</i>	207

Solicited Session 1
Environmental Assessment and Air Quality Indicators
Organizer: Antonella Plaia

F. Bruno, D. Cocchi <i>Interpreting air quality indices as random quantities</i>	213
D. Lee, C. Ferguson, E. M. Scott <i>Air quality indicators in health studies</i>	217
M. Ruggieri, A. Plaia, A. L. Bondi <i>Aggregated air pollution index: a new proposal</i>	221

Solicited Session 2
Variable Selection in Multivariate Analysis
Organizer: Gabriele Soffritti

N. Dean, A. E. Raftery <i>Latent class analysis variable selection</i>	227
C. Maugis, G. Celeux, M. Martin-Magniette <i>Variable selection in model-based clustering: a general variable role modeling</i>	231
L. Scrucca <i>A geometric approach to subset selection and sparse sufficient dimension reduction</i>	235

Solicited Session 3
Statistical Models for Electoral Flow Data
Organizer: Venera Tomaselli

J. M. Bernardo <i>Forecasting the final results in election day: a Bayesian analysis</i>	241
A. Forcina, G. M. Marchetti <i>The Brown and Payne model of voter transition revisited</i>	245
V. Tomaselli, R. D'Agata <i>Using multilevel models to analyze the context of electoral data</i>	249

Solicited Session 4
Partitioning Methods for Statistical Learning
Organizer: Massimo Aria

G. Boari, G. Cantaluppi <i>Further considerations on latent segmentation techniques for customer heterogeneity detection</i>	255
G. Giordano, R. Remmerswaal <i>Non-parametric regression model for a hierarchical data-structure a comparison with the classical approaches</i>	259

G. Soffritti, G. Galimberti <i>Detecting multiple cluster structures through model-based clustering methods</i>	263
--	-----

Solicited Session 5
Identification and Estimation of Causal Effects in the Presence of Complications
 Organizer: Leonardo Grilli

P. Frumento, B. Pacini <i>Causal inference with multivariate outcome: a simulation study</i>	269
A. Mattei <i>Identifying and estimating direct and indirect effects</i>	273
G. Roli, B. Pacini, <i>Evaluating the effects of financial aids to firms with nonignorably missing outcomes</i>	277

Solicited Session 6
Statistical Analysis of Social Networks of Enterprises
 Organizer: Francesco Palumbo

P. Ameli, F. Niccolini, F. Palumbo <i>Latent ties identification in inter-firms social networks</i>	283
M. R. D'Esposito, A. Del Monte, G. Giordano, M. P. Vitale <i>The analysis of structural characteristics of innovative networks</i>	287
S. Liverani, A. Petrucci <i>Spatial clustering of multivariate data using weighted MAX-SAT</i>	291

Solicited Session 7
Methodology for Evaluating the Effectiveness of the Italian University System
 Organizer: Luigi Fabbris

M. Bini <i>Evaluating the external effectiveness of the university education in Italy</i>	297
--	-----

G. Elias
*How to approach the evaluation of the effectiveness of higher education
in Italy* 301

L. Fabbris, G. Boccuzzo, M. C. Martini, M. Scioni
A participative process for the definition of a human capital indicator 305

Solicited Session 8
Advances in Data Collection for Surveys
Organizer: Pier Francesco Perri

G. Diana
Calibration estimators in randomized response models 311

I. Krumpal
*Estimating the prevalence of unsocial opinions via the randomized
response technique* 315

P. G. M. van der Heijden, U. Böckenholt
Randomized response, statistical disclosure control and e-commerce ... 319

Solicited Session 9
Functional Data Analysis
Organizer: Rosanna Verde

G. Costanzo, F. Dell'Accio, G. Trombetta
Prediction of an industrial kneading process via the adjustment curve ... 325

T. Di Battista, S. A. Gattone, A. De Sanctis
Dealing with FDA estimation methods 329

A. Goia
Functional linear regression: applications to time series prediction 333

Solicited Session 10
Econometric Analysis of Data on Market Behaviors
Organizer: Roberto Cellini

V. Adorno, C. Bernini, G. Pellegrini
Productivity, efficiency and public subsidies. An empirical evaluation 339

L. Arciero, A. Mercatanti <i>Evaluation of the effect of debit cards on cash demand by propensity score regressions in a principal stratification model</i>	343
R. Cellini, L. Lambertini, A. Sterlacchini <i>Managerial incentive and the firms' propensity to invest in product and process innovation</i>	347
M. Di Giacomo, M. Piacenza, G. Turati <i>The effectiveness of "flexible" specific taxation in stabilizing fuel prices. An empirical investigation on the Italian wholesale markets</i>	351

Solicited Session 11

From Market Segmentation to Different Customer Loyalties

Organizer: Mario Montinaro

P. Chirico, A. Lo Presti <i>Customer loyalty analysis in an heterogeneous market: a comparison between a priori segmentation and model-based segmentation</i>	357
P. Ferrari, A. Barbiero , G. Manzi <i>Missing data handling in setting up indicators from categorical data</i>	361
M. Montinaro, I. Sciascia <i>Probabilistic models for market segmentation and customer loyalty</i>	365
G. Nicolini, L. Dalla Valle <i>The problem of non random samples in customer satisfaction surveys</i> ..	369

Solicited Session 12

Markov and Hidden Markov Processes for Data Analysis and Simulation

Organizer: Isabella Morlini

F. Bartolucci, A. Farcomeni, F. Pennoni <i>Analysis of longitudinal data via latent Markov models and its extensions</i>	375
M. Costa, L. De Angelis <i>A dynamic analysis of stock markets through latent Markov models</i>	379

A. Maruotti, R. Rocci <i>A semiparametric approach to mixed nonhomogeneous hidden Markov models</i>	383
R. Paroli <i>Hidden Markov processes for image data analysis</i>	387

CONTRIBUTED PAPERS

G. Adelfio, M. Chiodi <i>Probabilistic forecast for Northern New Zealand seismic process: a kernel-based approach</i>	393
A. Altavilla, A. Mazza, M. Mucciardi <i>Differences between the settlement patterns of individual immigrant in the city of Catania</i>	397
A. Altavilla, M. Mondello <i>Mortality in environmental high risk area</i>	401
G. Amisano, R. Savona <i>Taming financial inaccessibility with Bayesian state space models</i>	405
G. Aretusi, L. Fontanella, L. Ippoliti, A. Merla <i>Supervised classification of thermal high-resolution IR images for the diagnosis of Raynaud's phenomenon</i>	409
S. Arima, V. Pecora, L. Tardella <i>Identifying epitopes in allergic reactions through hierarchical Bayesian modeling</i>	413
A. Balzanella, R. Verde, Y. Lechevallier <i>A new approach for clustering multiple streams of data</i>	417
S. Bande, S. Ghigo, R. Ignaccolo <i>Classification of municipalities with respect to air quality assessment</i>	421
L. Bertoli-Barsotti, A. Punzo <i>Modelling missingness in the Rasch paradigm</i>	425

G. Bianchi, F. M. L. Di Lascio, S. Giannerini, A. Manzari, A. Reale, G. Ruocco <i>Exploring copulas for the imputation of missing nonlinearly dependent data</i>	429
C. E. Bonafede, P. Cerchiello, P. Giudici <i>A semantic based Dirichlet compound multinomial model</i>	433
A. Brogini, D. Slanzi <i>Confident Bayesian networks: a non-parametric bootstrap approach</i>	437
F. Bruno, D. Cocchi, F. Greco <i>Functional cluster analysis for compositional data</i>	441
P. Brutti, S. Gubbiotti, V. Sambucini <i>A Bayesian two-stage design for phase II clinical trials with both efficacy and safety endpoints</i>	445
F. Campobasso, A. Fanizzi, M. Tarantini <i>A proposal for a fuzzy approach to clustering techniques</i>	449
M. Carpita, M. Manisera <i>On the nonlinearity of homogeneous ordinal variables</i>	453
A. Cossari <i>Using the bootstrap in the analysis of fractionated designs</i>	457
P. Cozzucoli <i>A control chart to monitor the overall defectiveness of the process</i>	461
F. Crippa, M. Marelli <i>Mixed-effects modeling for items and subjects in psycholinguistics: an application</i>	465
C. Cutugno <i>Model based clustering for solvency analysis</i>	469
G. D'Epifanio <i>A policy-planning based data-analysis in social agents's worthiness scaling</i>	473
F. Domma, S. Giordano <i>Dependence in stress-strength models with FGM copula and Burr III margins</i>	477

C. Drago, G. Scepi <i>Univariate and multivariate tools for visualizing financial time series</i>	481
S. Facchinetti, S. A. Osmetti <i>The Kolmogorov-Smirnov goodness of fit test for discrete extreme value distributions</i>	485
S. Figini <i>Local statistical models for variables selection</i>	489
M. Finocchiaro Castro, C. Guccio <i>Efficiency analysis of heritage authorities: a DEA-bootstrap approach</i>	493
P. Frumento <i>Using genetic algorithms in estimation of finite mixture models</i>	497
M. Gallo, D. Fiorenzo <i>A strategy to evaluate the satisfaction of different passenger groups</i>	501
R. Gargano, A. Alibrandi <i>Modeling resistin levels data by mixture of regression</i>	505
A. Geyer-Schulz, B. Hoser <i>Financial accounting and eigensystem analysis</i>	509
P. Giudici, E. Raffinetti <i>On dependence measures in a multivariate context</i>	513
S. Golia <i>The impact of uniform and nonuniform differential item functioning on Rasch measure</i>	517
M. Gnaldi, M. Ranalli <i>Ranking the Italian universities: a sensitivity analysis of research performance composite indicators</i>	521
M. Grassia, D. Nappo <i>An algorithm for the treatment of qualitative variables in a SEM model and its validation</i>	525
F. Greselin, S. Ingrassia, A. Punzo <i>Multivariate tests for patterned covariance matrices</i>	529

R. Guido, M. Misuraca, F. Vocaturo <i>Document classification in public administration: an automatic strategy for digital protocol</i>	533
Y. Henchiri, S. Benammou, T. Hamza <i>Using support vector machine approach for forecasting the failure of the Tunisian companies</i>	537
A. Lourme, C. Biernacki <i>Joint clustering of two samples from multiple origins by equalizing their entropy</i>	541
D. Marasini, P. Quatto <i>Control sample and association</i>	545
P. Mariani, E. Zavarrone <i>Satisfaction, loyalty and WOM in dental care sector</i>	549
P. Mariani, E. Zavarrone <i>The measure of economic re-evaluation: a coefficient based on conjoint analysis</i>	553
M. Marozzi <i>Some notes on indicator variable reduction</i>	557
M. Matteucci, S. Mignani, R. Ricci <i>Latent variable models to evaluate the final exam in the Italian lower secondary school</i>	561
M. Matteucci, B. Veldkamp <i>The inclusion of prior information in CAT administration</i>	565
A. Mazza, A. Punzo <i>Graduation of demographic data using discrete beta kernel estimates</i> ...	569
I. Mdimagh, S. Benammou <i>Using partial least squares regression in lifetime analysis</i>	573
F. Mealli, C. Rampichini <i>Exploiting nonlinear difference-in-difference assumptions in a regression discontinuity design</i>	577

S. C. Minotti, G. Spedicato <i>A statistical R package for cluster-weighted modeling</i>	581
I. Morlini, S. Zani <i>New weighted similarity indexes</i>	585
A. M. Oliveri, A. M. Parroco <i>CAPI versus PAPI and the production of some non-sampling errors</i>	589
A. Pastore, S. F. Tonellato <i>A class of indexes for comparing model-based clustering solutions</i>	593
I. Poli, D. Slanzi, L. Villanova, D. De Lucrezia, G. Minervini, F. Polticelli <i>Non natural protein structure identification</i>	597
M. Porcu, I. Sulis, N. Tedesco <i>Evaluating lecturer's capability over time. Some evidence from surveys on university course quality</i>	601
E. Rocco <i>Kernel-type smoothing methods in the nonresponse context</i>	605
E. Romano, A. Balzanella, R. Verde <i>A clusterwise regression strategy for spatio-functional data</i>	609
E. Romano, V. Cozza <i>Clustering spatio-functional data: a method based on a nonparametric variogram estimation for functional data</i>	613
A. A. Romano, G. Scandurra <i>Impact of exogenous shocks on oil product market prices</i>	617
G. Schoier <i>On the problem of variable reduction in a customer satisfaction context</i>	621
B. Simonetti, L. D'Ambra, P. Amenta <i>New developments in ordinal non symmetrical correspondence analysis</i>	625
A. Tarsitano <i>A weighted distance for short time series</i>	629

F. Torti, D. Perrotta <i>The test size of the forward search for regression outliers</i>	633
M. Tsuji, S. Konsha, T. Shimokawa <i>A presentational approach to clustering the regional structural analysis with time</i>	637
V. A. Tutore, A. D'Ambrosio <i>Three-way data analysis by tree-based partitioning</i>	641
M. Vezzoli, P. Zuccolotto <i>Variable importance measurement within hierarchical data</i>	645
A. Zarraga, B. Goitisoló <i>Correspondence analysis of surveys with multiple response questions</i> ..	649
Authors Index	653