



SIS - CLADAG



# CLADAG 2015

10° Scientific Meeting of the Classification and Data Analysis  
Group of the Italian Statistical Society

Flamingo Resort, Santa Margherita di Pula,  
October 8-10, 2015

## BOOK OF ABSTRACTS

### Editors:

Francesco Mola, Claudio Conversano  
CUEC Editrice, Cagliari



CUEC  
editrice

ISBN: 978 88 8467 749 9



Univeristà degli Studi  
di Cagliari



Fondazione  
Banco di Sardegna

## CUEC EDITRICE

by Sardegna Novamedia Soc. Coop.

Via Basilicata n. 57/59

09127 Cagliari,

ITALY

Tel. & Fax +39 070 271573

[www.cuec.eu](http://www.cuec.eu)

[info@cuec.eu](mailto:info@cuec.eu)

ISBN: 978-88-8467-749-9

First Edition CUEC © 2015

## PREFACE

CLADAG 2015, the 10th Scientific Meeting of the Classification and Data Analysis Group of the Italian Statistical Society (SIS), will be held in Santa Margherita di Pula, Cagliari, Italy, from October 8th to October 10th 2015. The local organizer is the Department of Business and Economics of the University of Cagliari.

CLADAG 2015 will take place under the auspices of the International Federation of Classification Societies (IFCS) and of the Italian Statistical Society (SIS). It promotes advanced methodological research in multivariate statistics with a special vocation in Data Analysis and Classification. CLADAG supports the interchange of ideas in these fields of research, including the dissemination of concepts, numerical methods, algorithms, computational and applied results. It will also benefit of the support of Fondazione Banco di Sardegna.

CLADAG is a member of the International Federation of Classification Societies (IFCS). Among its activities, CLADAG organizes a biennial scientific meeting, schools related to classification and data analysis, publishes a newsletter, and cooperates with other member societies of the IFCS to the organization of their conferences.

The scientific program comprises three Keynote Lectures, an Invited Session, 10 Specialized Sessions, 15 Solicited Sessions and 15 Contributed Sessions. All the Specialized and Solicited Sessions have been promoted by the members of the Scientific Program Committee. The organizers wish to thank them for their cooperation in contributing to the success of CLADAG 2015.

The Book of Abstracts contains short papers of all the presentations scheduled in the conference program. It is organized according to type of session/lecture: Keynote Lectures, Specialized Sessions, Solicited Sessions and Contributed Sessions.

The editors would like to express their gratitude to the Rector of the University of Cagliari, the Director of the Department of Business and Economics and to all the statisticians working in the Department of Business and Economics for their enthusiasm in supporting the organization of this event from the very beginning, as well as to all people who worked hard to make it a success. Special thanks go to Dr. Massimo Cannas, Dr. Luca Frigau and Dr. Farideh Tavazoee for their editorial support

Last but not least, we thank all authors and participants, without whom the conference would not have been possible.

Cagliari, October 8 2015.

*Francesco Mola,  
Claudio Conversano*

## CONFERENCE THEMES

The 10th Meeting is orientated towards all topics related to data analysis, classification, multivariate and computational statistics. Submission of papers addressing these topics in both methodological and practical perspective has been encouraged by the members of the Scientific Program Committee.

The list of topics includes, but is not limited to, the following:

### **A Classification Theory**

Bayesian Classification Biplots Clustering models Consensus of Classifications Correspondence Analysis Discrimination and Classification Factor Analysis and Dimension Reduction Methods Fuzzy Methods Genetic Algorithms Hierarchical Classification Multidimensional Scaling Multiway Scaling Multiway Methods Neural Networks for Classification Non Hierarchical Classification Similarities and Dissimilarities Software algorithms for classification Unfolding and Related Scaling Methods

### **B Data Analysis**

Bayesian data Analysis Big data analysis- Categorical Data Analysis Covariance Structure Analysis Data Mining Data Science Data Visualization Decision Trees Functional data analysis Mixture and Latent Class Models Multilevel data Analysis Non Linear Data Analysis Nonparametric and Semiparametric Regression Partial Least Squares Pattern recognition Robustness and Data Diagnostics Social networks- Software algorithms for multivariate analysis Spatial Data Analysis Symbolic Data Analysis.

## COMMITTEES

### Scientific Program Committee

**Chair:** Paolo Giudici (*University of Pavia*)

#### Members

Giuseppe Bove (*University of Roma Tre*)  
Daniela Calo (*University of Bologna*)  
Agostino Di Ciaccio (*University of Roma La Sapienza*)  
Vincenzo Esposito Vinzi (*ESSEC, France*)  
Francesca Greselin (*University of Milano Bicocca*)  
Francesco Mola (*University of Cagliari*)  
Francesco Palumbo (*University of Naples Federico II*)  
Carla Rampichini (*University of Firenze*)  
Giancarlo Ragozini (*University of Naples Federico II*)  
Fabrizio Ruggeri (*CNR IMATI, Milan*)  
Silvia Salini (*University of Milan*)  
Adalbert F.X. Wilhelm (*Jacobs University Bremen, Germany*)

### Local Organizing Committee

**Chair:** Francesco Mola (*University of Cagliari, Italy*).

**Members:** Stefano Cabras, Massimo Cannas, Claudio Conversano, Luca Frigau, Monica Musio, Mariano Porcu, Luisa Salaris, Isabella Sulis, Nicola Tedesco.

## PARTICIPATING ORGANIZATIONS



International Federation of Classification Societies (IFCS)



Società Italiana di Statistica (SIS)



SIS - CLADAG (Classification and Data Analysis Group of the Italian Statistical Society)



Università degli Studi di Cagliari



Fondazione Banco di Sardegna



# Table of Contents

## Keynote Lectures

MINING KEY NETWORKS . . . . .	21
<i>David Banks</i>	
VARIABLE SELECTION FOR MODEL-BASED CLUSTERING OF CATEGORICAL DATA . . . . .	22
<i>Brendan Murphy</i>	
EIGENVALUES IN MIXTURE MODELING: GEOMETRIC, ROBUSTNESS AND COMPUTATIONAL ISSUES . . . . .	23
<i>Salvatore Ingrassia</i>	

## Specialized sessions

### Robust methods for the analysis of Economic (Big) data

*Organizer and Chair: Silvia Salini*

FAST AND ROBUST SEEMINGLY UNRELATED REGRESSION . . . . .	25
<i>Mia Hubert, Tim Verdonck and Ozlem Yorulmaz</i>	
APPLICATION TO THE DETECTION OF CUSTOMS FRAUD OF THE GOODNESS-OF-FIT TESTING FOR THE NEWCOMB-BENFORD LAW . . . . .	30
<i>Lucio Barabesi, Andrea Cerasa, Andrea Cerioli and Domenico Perrotta</i>	
MONITORING THE ROBUST ANALYSIS OF A SINGLE MULTIVARIATE SAMPLE . . . . .	34
<i>Marco Riani, Anthony C. Atkinson and Andrea Cerioli</i>	

### Bayesian nonparametric clustering

*Organizer: Fabrizio Ruggeri; Chair: Renata Rotondi*

A BAYSIAN NONPARAMETRIC APPROACH TO MODEL ASSOCIATION BETWEEN CLUSTERS OF SNPS AND DISEASE RESPONSES . . . . .	39
<i>Raffaele Argiento, Alessandra Guglielmi, Chuhsing Kate Hsiao, Fabrizio Ruggeri and Charlotte Wang</i>	
A BAYESIAN NONPARAMETRIC MODEL FOR CLUSTERING AND BORROWING INFORMATION . . . . .	43
<i>Antonio Lijoi, Bernardo Nipoti and Igor Prünster</i>	
SEQUENTIAL CLUSTERING BASED ON DIRICHLET PROCESS PRIORS . . . . .	47
<i>Roberto Casarin, Andrea Pastore and Stefano F. Tonellato</i>	



## Causal Inference with Complex Data Structures

*Organizer and Chair: Alessandra Mattei*

SHORT TERM IMPACT OF PM10 EXPOSURE ON MORTALITY: A PROPENSITY SCORE APPROACH . . . . .	52
--	----

*Michela Baccini, Alessandra Mattei and Fabrizia Mealli*

IDENTIFICATION AND ESTIMATION OF CAUSAL MECHANISMS IN CLUSTERED ENCOURAGEMENT DESIGNS: DISENTANGLING BED NETS USING BAYESIAN PRINCIPAL STRATIFICATION . . . . .	54
---	----

*Laura Forastiere, Fabrizia Mealli and Tyler VanderWeele*

THE EFFECTS OF A DROPOUT PREVENTION PROGRAM ON SECONDARY STUDENTS' OUTCOMES . . . . .	56
---	----

*Enrico Conti , Silvia Duranti, Alessandra Mattei, Fabrizia Mealli and Nicola Sciclone*

## Clustering in Time Series

*Organizer and Chair: Michele La Rocca*

PROBABILISTIC BOOSTED-ORIENTED CLUSTERING OF TIME SERIES . . . . .	61
--	----

*Antonio D'Ambrosio, Gianluca Frasso, Carmela Iorio and Roberta Siciliano*

COPULA-BASED FUZZY CLUSTERING OF TIME SERIES . . . . .	65
--	----

*Pierpaolo D'Urso, Marta Disegna and Fabrizio Durante*

COMPARING MULTI-STEP AHEAD FORECASTING FUNCTIONS FOR TIME SERIES CLUSTERING . . . . .	69
---	----

*Marcella Corduas and Giancarlo Ragozini*

## Multiway Analysis

*Organizer and Chair: Giuseppe Bove*

(INTERACTIVE) VISUALISATION OF THREWAY DATA . . . . .	74
---	----

*Casper J. Albers and John C. Gower*

ROBUST FUZZY CLUSTERING OF MULTIVARIATE TIME TRAJECTORIES . . . . .	78
---	----

*Pierpaolo D'Urso and Riccardo Massari*

ESTIMATION PROCEDURES FOR AVOIDING DEGENERATE SOLUTIONS IN CANDECOMP/PARAFAC . . . . .	82
--	----

*Paolo Giordani*

**Big Data Analysis**

*Organizer and Chair: Donato Malerba*

TOWARDS A STATISTICAL FRAMEWORK FOR ATTRIBUTE COMPARISON IN VERY LARGE RELATIONAL DATABASES . . . . . 83

*Cesare Alippi, Elisa Quintarelli, Manuel Roveri and Letizia Tanca*

MINING BIG DATA WITH HIGH PERFORMANCE COMPUTING SOLUTIONS . . . . . 91

*Fabrizio Angiulli, Stefano Basta, Stefano Lodi, Gianluca Moro and Claudio Sartori*

ENHANCING BIG DATA EXPLORATION WITH FACETED BROWSING . . . . . 95

*Sonia Bergamaschi, Giovanni Simonini and Song Zhu*

**New Methodologies for Composite Indicators**

*Organizer and Chair: Agostino Di Ciaccio*

ADVANCES IN COMPOSITE-BASED PATH MODELING FOR SYNTHETIC INDICATORS 100

*Vincenzo Esposito Vinzi, Laura Trinchera and Giorgio Russolillo*

COMPOSITE INDICATORS MODELING . . . . . 102

*Maurizio Vichi*

MEASURING THE IMPORTANCE OF VARIABLES IN COMPOSITE INDICATORS . . . 104

*William Becker, Michaela Saisana, Paolo Paruolo and Andrea Saltelli*

**Cluster analysis software and validation**

*Organizer and Chair: Christian Hennig*

ADAPTIVE CHOICE OF INPUT PARAMETERS IN ROBUST CLUSTERING . . . . . 109

*Luis A. García-Escudero and Augustin Mayo-Iscar*

ROBUST MODEL-BASED CLUSTERING WITH COVARIANCE MATRIX CONSTRAINTS 113

*Pietro Coretto and Christian Hennig*

FLEXIBLE IMPLEMENTATION OF RESAMPLING SCHEMES FOR CLUSTER VALIDATION 117

*Friedrich Leisch*

**Selecting a mixture model with a clustering focus**

*Organizer and Chair: Gilles Celeux*

CLUSTERING IN FINITE MIXTURES USING AN INTEGRATED COMPLETED LIKELIHOOD CRITERION . . . . . 122

*Marco Bertolotti, Nial Friel and Riccardo Rastelli*

ESTIMATION AND MODEL SELECTION FOR MODEL-BASED CLUSTERING WITH THE CONDITIONAL CLASSIFICATION LIKELIHOOD . . . . . 126

*Jean-Patrick Baudry*

ON THE DIFFERENT WAYS TO COMPUTE THE INTEGRATED COMPLETED LIKELIHOOD CRITERION . . . . .	130
<i>Gilles Celeux</i>	
<b>Exploring relationships between blocks of variables</b>	
<i>Organizer and Chair: Giorgio Russolillo</i>	
WEIGHTED MULTIBLOCK CLUSTERING . . . . .	135
<i>Ndéye Niang and Mory Ouattara</i>	
THEMATIC MODEL EXPLORATION THROUGH MULTIPLE CO-STRUCTURE MAXIMISATION: METHOD AND SOFTWARE . . . . .	139
<i>Xavier Bry and Thomas Verron</i>	
A NEW COMPONENT-BASED APPROACH OF REGULARISATION FOR MULTIVARIATE GENERALISED LINEAR REGRESSION . . . . .	144
<i>Catherine Trottier, Xavier Bry, Frederic Mortier and Guillaume Cornu</i>	
<b>Solicited Sessions</b>	
<b>Advances in Density-based clustering</b>	
<i>Organizer and Chair: Francesca Greselin</i>	
A NONPARAMETRIC CLUSTERING METHOD FOR IMAGE SEGMENTATION . . . . .	150
<i>Giovanna Menardi</i>	
ROBUST CLUSTERING FOR HETEROGENOUS SKEW DATA . . . . .	154
<i>Luis A. García-Escudero, Francesca Greselin and Agustin Mayo-Iscar</i>	
REGULARIZING FINITE MIXTURES OF GAUSSIAN DISTRIBUTIONS . . . . .	154
<i>Bettina Grün and Gertraud Malsiner-Walli</i>	
<b>Latent variable models for longitudinal data - Part I</b>	
<i>Organizer and Chair: Silvia Bacci</i>	
A JOINT MODEL FOR LONGITUDINAL AND SURVIVAL DATA BASED ON AN AR(1) LATENT PROCESS . . . . .	163
<i>Silvia Bacci, Francesco Bartolucci and Silvia Pandolfi</i>	
FINITE MIXTURE MODELS FOR MIXED DATA: EM ALGORITHMS AND PARAFAC REPRESENTATIONS . . . . .	167
<i>Marco Alfó and Paolo Giordani</i>	
ON THE USE OF THE CONTAMINATED GAUSSIAN DISTRIBUTION IN HIDDEN MARKOV MODELS FOR LONGITUDINAL DATA . . . . .	171
<i>Antonio Punzo and Antonello Maruotti</i>	
<b>Latent variable models for longitudinal data - Part II</b>	
<i>Organizer and Chair: Francesco Bartolucci</i>	
A HIDDEN MARKOV APPROACH TO THE ANALYSIS OF INCOMPLETE MULTIVARIATE LONGITUDINAL DATA . . . . .	177
<i>Francesco Lagona</i>	
LATENT MARKOV AND GROWTH MIXTURE MODELS: A COMPARISON . . . . .	181

*Fulvia Pennoni and Isabella Romeo*

LATENT WORTHS AND LONGITUDINAL PAIRED COMPARISONS - A MARKOV MODEL OF DEPENDENCE . . . . .	185
--	-----

*Brian Francis, Alexandra Grand and Regina Dittrich*

**Multivariate data analysis in environmental sciences**

*Organizer: Fabrizio Ruggeri; Chair: Raffaele Argiento*

MULTIVARIATE DOWNSCALING FOR NON-GAUSSIAN DATA . . . . .	191
--	-----

*Daniela Cocchi, Lucia Paci and Carlo Trivisano*

PRELIMINARY RESULTS ON TAPERING MULTIVARIATE SPATIO TEMPORAL MODELS FOR EXPOSURE TO AIRBORNE MULTIPOLLUTANTS IN EUROPE . . . . .	195
--	-----

*Alessandro Fassó, Francesco Finazzi and Ferdinand Ndongo*

CLUSTERING MACROSEISMIC FIELDS BY STATISTICAL DATA DEPTH FUNCTIONS	199
--	-----

*Claudio Agostinelli, Renata Rotondi and Elisa Varini*

**Advanced models for tourism analysis**

*Organizer and Chair: Stefania Mignani*

ANALYSING TERRITORIAL HETEROGENEITY IN TOURISTS'SATISFACTION TOWARDS ITALIAN DESTINATIONS . . . . .	204
---	-----

*Cristina Bernini, Augusto Cerqua and Guido Pellegrini*

MICRO-ECONOMIC DETERMINANTS OF TOURIST EXPENDITURE: A QUANTILE REGRESSION APPROACH . . . . .	208
--	-----

*Emanuela Marrocu, Raffaele Paci and Andrea Zara*

INEQUALITIES AND TOURISM CONSUMPTION BEHAVIOUR: A MIXTURE MODEL ANALYSIS . . . . .	212
--	-----

*Cristina Bernini, Maria Francesca Cracolici and Cinzia Viroli*

**Bayesian Networks and Graphical Models in Socio-Economic Sciences**

*Organizer and Chair: Paola Vicard*

BAYESIAN NETWORKS FOR FIRM PERFORMANCE EVALUATION . . . . .	217
---	-----

*Maria E. De Giuli, Pietro Gottardo, Anna M. Moisello and Claudia Tarantola*

GRAPHICAL MODEL USING COPULAS FOR MEASUREMENT ERROR MODELING . .	221
--	-----

*Daniela Marella and Paola Vicard*

**Time Series in Clustering**

*Organizer and Chair: Michele La Rocca*

PARSIMONIOUS CLUSTERING OF TIME SERIES . . . . .	226
--	-----

*Carmela Iorio, Antonio D'Ambrosio, Gianluca Frasso and Roberta Siciliano*

DYNAMIC TIME WARPING-BASED FUZZY CLUSTERING FOR SPATIAL TIME SERIES	230
---	-----

*Pierpaolo D'Urso, Marta Disegna and Riccardo Massari*

PERIODICAL FEATURE BASED TIME SERIES CLUSTERING . . . . .	234
---	-----

*Francesco Giordano, Michele La Rocca and Maria Lucia Parrella*

## Big Data Analysis

*Organizer and Chair: Donato Malerba*

INTERACTIVE MACHINE LEARNING WITH R . . . . . 239

*Giorgio Maria Di Nunzio*

WORKLOAD ESTIMATION FOR A CALL CENTER . . . . . 243

*Pierluigi Riva and Ruggiero Scommegna*

PREDICTION IN OLIVE OIL TRADE USING REGRESSION MODELS ON TEMPORAL  
DATA NETWORK . . . . . 245

*Corrado Loglisci , Umberto Medicamento and Arturo Casieri*

## Advances in Ordinal and Preference Data

*Organizer and Chair: Antonio D'Ambrosio*

MEASURING CONSENSUS IN THE SETTING OF NON-UNIFORM QUALITATIVE SCALES 250

*José L. García-Lapresta and David Pérez-Román*

ACCURATE ALGORITHMS FOR CONSENSUS RANKING DETECTION . . . . . 255

*Giulio Mazzeo, Antonio D'Ambrosio and Roberta Siciliano*

LOGISTIC REGRESSION TREES FOR ORDINAL AND PREFERENCE DATA . . . . . 259

*Thomas Rusch, Achim Zeileis and Kurt Hornik*

## Case studies in data science from Ligurian companies

*Organizer and Chair: Delio Panaro*

STATISTICAL METHODS FOR THE ANALYSIS OF OSTREOPSIS OVATA BLOOM EVENTS  
FROM METEO-MARINE DATA . . . . . 262

*Ennio Ottaviani, Valentina Asnaghi, Mariachiara Chiantore, Andrea Pedroncini  
and Rosella Bertolotto*

DATA MINING FOR OPTIMAL GAMBLING . . . . . 266

*Gabriele Torre and Fabrizio Malfanti*

A FRAUD DETECTION ALGORITHM FOR ONLINE BANKING . . . . . 271

*Delio Panaro, Eva Riccomagno and Fabrizio Malfanti*

DOES DIRECTORS' BACKGROUND MATTER? FIRM PERFORMANCE, BOARD FEA-  
TURES AND FINANCIAL REPORTING RELIABILITY . . . . . 275

*Delio Panaro, Silvia Ferramosca and Sara Trucco*

**Modeling ordinal data**

*Organizer and Chair: Maurizio Carpita*

POSTERIOR PREDICTIVE MODEL CHECKS FOR ASSESSING THE GOODNESS OF FIT OF BAYESIAN MULTIDIMENSIONAL IRT MODELS . . . . . 280

*Mariagiulia Matteucci and Stefania Mignani*

INTERNATIONAL TOURISM IN ITALY: A BAYESIAN NETWORK APPROACH . . . . . 284

*Federica Cugnata and Giovanni Perucca*

CLUSTERING UPPER LEVEL UNITS IN MULTILEVEL MODELS FOR ORDINAL DATA 288

*Leonardo Grilli, Agnese Panzera and Carla Rampichini*

**Functional data analysis for environmental data**

*Organizer and Chair: Tonio Di Battista*

CLUSTERING SPATIALLY DEPENDENT FUNCTIONAL DATA: A METHOD BASED ON THE CONCEPT OF SPATIAL DISPERSION FUNCTION OF A CURVE . . . . . 292

*Elvira Romano, Antonio Balzanella and Rosanna Verde*

TWO CASE STUDIES ON OBJECT ORIENTED SPATIAL STATISTICS . . . . . 296

*Piercesare Secchi, Simone Vantini and Valeria Vitelli*

INFERENCE ON FUNCTIONAL BIODIVERSITY TOOLS . . . . . 298

*Tonio Di Battista, Francesca Fortuna and Fabrizio Maturo*

**Advances in quantile regression**

*Organizer and Chair: Cristina Davino*

M-QUANTILE REGRESSION: DIAGNOSTICS AND PARAMETRIC REPRESENTATION OF THE MODEL . . . . . 303

*Annamaria Bianchi, Enrico Fabrizi, Nicola Salvati and Nikos Tzavidis*

QUANTILE REGRESSION: A BAYESIAN ROBUST APPROACH . . . . . 307

*Marco Bottone, Mauro Bernardi and Lea Petrella*

A COMPARISON AMONG ESTIMATORS FOR LINEAR REGRESSION METHODS . . . . 311

*Marilena Furno and Domenico Vistocco*

HANDLING HETEROGENEITY AMONG UNITS IN QUANTILE REGRESSION . . . . . 315

*Cristina Davino and Domenico Vistocco*

**Directional Data**

*Organizer and Chair: Giovanni C. Porzio*

SMALL BIASED CIRCULAR DENSITY ESTIMATION . . . . . 320

*Marco Di Marzio, Stefania Fensore, Agnese Panzera, and Charles C. Taylor*

A DEPTH-BASED CLASSIFIER FOR CIRCULAR DATA . . . . . 324

*Giuseppe Pandolfo*

NONPARAMETRIC ESTIMATES OF THE MODE FOR DIRECTIONAL DATA . . . . . 328  
*Thomas Kirschstein, Steffen Liebscher, Giovanni C. Porzio and Giancarlo Ragozini*

**Recent developments in statistical analysis of network data**

*Organizer and Chair: Domenico De Stefano*

GAME THEORY AND NETWORK MODELS FOR THE RECONSTRUCTION OF AR-  
CHAEOLOGICAL NETWORKS . . . . . 331  
*Viviana Amati and Ulrik Brandes*

A MODEL FOR CLUSTERING A SPATIAL NETWORK WITH APPLICATION TO LOCAL  
LABOUR SYSTEM IDENTIFICATION . . . . . 335  
*Francesco Pauli, Nicola Torelli and Susanna Zaccarin*

ON THE SAMPLING DISTRIBUTIONS OF THE ML ESTIMATORS IN NETWORK EF-  
FECT MODELS . . . . . 339  
*Michele La Rocca, Giovanni C. Porzio, Maria Prosperina Vitale and Patrick Dor-  
eian*

CORRESPONDENCE ANALYSIS WITH DOUBLING FOR TWO-MODE VALUED NET-  
WORKS . . . . . 343  
*Giancarlo Ragozini, Domenico De Stefano and Daniela D'Ambrosio*

**Current challenges in clustering and classification of biomedical data**

*Organizer and Chair: Adalbert F.X. Wilhelm*

SEMANTIC MULTI CLASSIFIER SYSTEMS FOR THE DETECTION OF AGING RELATED  
PROCESSES . . . . . 348  
*Hans A. Kestler, Ludwig Lausser, Lyn-Rouven Schirra, Florian Schmid*

EMOTION RECOGNITION IN HUMAN COMPUTER INTERACTION USING MULTIPLE  
CLASSIFIER SYSTEMS . . . . . 349  
*Friedhelm Schwenker*

ENSEMBLE OF SELECTED CLASSIFIERS . . . . . 352  
*Berthold Lausen, Asma Gul, Zardad Khan and Osama Mahmoud*

**Contributed papers**

A GENERALIZED DISTANCE FOR INFERENCE IN FUNCTIONAL DATA . . . . . 354  
*Andrea Ghiglietti and Anna M. Paganoni*

LONG GAPS IN MULTIVARIATE SPATIO-TEMPORAL DATA: AN APPROACH BASED  
ON FUNCTIONAL DATA ANALYSIS . . . . . 359  
*Mariantonietta Ruggieri, Antonella Plaia and Francesca Di Salvo*

EFFECTS ON CURVE CLUSTERING OF DIFFERENT TRANSFORMATIONS OF CHRONO-  
LOGICAL TEXTUAL DATA . . . . . 363  
*Matilde Trevisani and Arjuna Tuzzi*

A NOTE ON THE RELIABILITY OF A CLASSIFIER . . . . . 366  
*Luca Frigau*

ROBUSTIFIED CLASSIFICATION OF MULTIVARIATE FUNCTIONAL DATA . . . . .	370
<i>Francesca Ieva and Anna M. Paganoni</i>	
SIZE CONTROL OF ROBUST REGRESSION ESTIMATORS . . . . .	374
<i>Silvia Salini, Andrea Cerioli, Fabrizio Laurini and Marco Riani</i>	
THE MOVEMENTS OF EMOTIONS: AN EXPLORATORY CLASSIFICATION ON AF- FECTIVE MOVEMENT DATA . . . . .	378
<i>Pasquale Dente, Arvid Kappas and Adalbert F. X. Wilhelm</i>	
ELECTRE TRI-MACHINE LEARNING APPROACH TO THE RECORD LINKAGE PROBLEM	382
<i>Valentina Minnetti and Renato De Leone</i>	
QUALITY OF CLASSIFICATION APPROACHES FOR THE QUANTITATIVE ANALYSIS OF INTERNATIONAL CONFLICT . . . . .	387
<i>Adalbert F.X. Wilhelm</i>	
THE RTCLUST PROCEDURE FOR ROBUST CLUSTERING . . . . .	391
<i>Francesco Dotto, Alessio Farcomeni, Luis Angel García-Escudero and Agustín Mayo- Iscar</i>	
WHAT ARE THE TRUE CLUSTERS? . . . . .	396
<i>Christian Hennig</i>	
A NOVEL MODEL-BASED CLUSTERING APPROACH FOR MASSIVE DATASETS OF SPATIALLY REGISTERED TIME SERIES. WITH APPLICATION TO SEA SURFACE TEMPERATURE REMOTE SENSING DATA . . . . .	399
<i>Francesco Finazzi and Marian Scott</i>	
BIG DATA CLASSIFICATION: SIMULATIONS IN THE MANY FEATURES CASE . . . .	403
<i>Claus Weihs</i>	
FROM BIG DATA TO INFORMATION: STATISTICAL ISSUES THROUGH EXAMPLES .	407
<i>Silvia Biffignandi and Serena Signorelli</i>	
BIG DATA MEET PHARMACEUTICAL INDUSTRY: AN APPLICATION ON SOCIAL MEDIA DATA . . . . .	411
<i>Caterina Liberati and Paolo Mariani</i>	
DEFINING THE SUBJECTS DISTANCE IN HIERARCHICAL CLUSTER ANALYSIS BY COPULA APPROACH . . . . .	416
<i>Andrea Bonanomi, Marta Nai Ruscone and Silvia Angela Osmetti</i>	



SUPERVISED CLASSIFICATION OF DEFECTIVE CRANKSHAFTS BY IMAGE ANALYSIS	420
<i>Beatriz Remeseiro, Javier Tarrío-Saavedra, Mario Francisco-Fernández, Manuel G. Penedo, Salvador Naya and Ricardo Cao</i>	
ARCHETYPAL ANALYSIS FOR DATA-DRIVEN PROTOTYPE IDENTIFICATION . . . . .	424
<i>Giancarlo Ragozini, Francesco Palumbo and Maria R. D'Esposito</i>	
PRINCIPAL COMPONENT ANALYSIS OF COMPLEX DATA AND APPLICATION TO CLIMATOLOGY . . . . .	428
<i>Sergio Camiz and Silvia Creta</i>	
SPARSE EXPLORATORY MULTIDIMENSIONAL IRT MODELS . . . . .	432
<i>Lara Fontanella, Sara Fontanella, Pasquale Valentini, Nickolay Trendafilov</i>	
ITERATIVE FACTOR CLUSTERING FOR CATEGORICAL DATA RECONSIDERED . . . . .	437
<i>Alfonso Iodice D'Enza, Angelos Markos and Francesco Palumbo</i>	
TESTING ANTIPODAL SYMMETRY OF CIRCULAR DATA . . . . .	442
<i>Giovanni Casale, Giuseppe Pandolfo and Giovanni C. Porzio</i>	
HOW TO DEFINE DEVIANCE RESIDUALS IN MULTINOMIAL REGRESSION . . . . .	446
<i>Giovanni Romeo, Mariangela Sciandra and Marcello Chiodi</i>	
DIAGNOSTIC TOOLS FOR GAMLSS FITTED OBJECTS . . . . .	451
<i>Andrea Marletta and Mariangela Sciandra</i>	
BAYESIAN REGRESSION ANALYSIS WITH LINKED AND DUPLICATED DATA . . . . .	455
<i>Andrea Tancredi, Rebecca Steorts and Brunero Liseo</i>	
A SEMI-PARAMETRIC FAY-HERRIOT-TYPE MODEL WITH UNKNOWN SAMPLING VARIANCES . . . . .	460
<i>Silvia Polettini</i>	
POSTERIOR DISTRIBUTIONS FROM OPTIMALLY B-ROBUST ESTIMATING FUNCTIONS AND APPROXIMATE BAYESIAN COMPUTATION . . . . .	464
<i>Ivan Luciano Danesi, Fabio Piacenza, Erlis Ruli and Laura Ventura</i>	
MCA BASED COMMUNITY DETECTION . . . . .	468
<i>Carlo Drago</i>	
CLASSIFYING SOCIAL ROLES BY NETWORK STRUCTURES . . . . .	472
<i>Simona Gozzo and Venera Tomaselli</i>	
A MULTILEVEL HECKMAN MODEL TO INVESTIGATE FINANCIAL ASSETS AMONG OLD PEOPLE IN EUROPE . . . . .	476
<i>Omar Paccagnella and Chiara Dal Bianco</i>	
OPTIMAL PRICING USING BAYESIAN SEMIPARAMETRIC PRICE RESPONSE MODELS	480
<i>Winfried J. Steiner, Anett Weber, Stefan Lang and Peter Wechselberger</i>	

MONETARY TRANSMISSION MODELS FOR BANKING INTEREST RATES . . . . .	484
<i>Laura Parisi, Paolo Giudici, Igor Gianfrancesco and Camillo Giliberto</i>	
ESTIMATING THE EFFECT OF PRENATAL CARE ON BIRTH OUTCOMES . . . . .	490
<i>Emiliano Sironi and Massimo Cannas</i>	
RECURSIVE PARTITIONING: AN APPROACH BASED ON THE WEIGHTED KEMENY DISTANCE . . . . .	494
<i>Mariangela Sciandra, Antonella Plaia and Veronica Picone</i>	
WHY TO STUDY ABROAD? AN EXAMPLE OF CLUSTERING . . . . .	498
<i>Valeria Caviezel and Anna M. Falzoni</i>	
A GRAPHICAL COPULA-BASED TOOL FOR DETECTING TAIL DEPENDENCE . . . . .	502
<i>Roberta Pappadà, Fabrizio Durante and Nicola Torelli</i>	
CLASSIFICATION MODELS AS TOOLS OF BANKRUPTCY PREDICTION - POLISH EXPERIENCE . . . . .	506
<i>Józef Pocięcha, Barbara Pawełek, Mateusz Baryła and Sabina Augustyn</i>	
THE RELATIONSHIP BETWEEN INDIVIDUAL PRICE RESPONSE OF BEER CONSUMERS AND THEIR DEMOGRAPHIC/PSYCHOGRAPHIC CHARACTERISTICS . . . . .	510
<i>Friederike Paetz</i>	
THE ENSEMBLE CONCEPTUAL CLUSTERING OF SYMBOLIC DATA FOR CUSTOMER LOYALTY ANALYSIS . . . . .	514
<i>Marcin Pełka</i>	
INSERT HERE CONSUMERS' PERCEPTIONS OF CORPORATE SOCIAL RESPONSIBILITIES AND WILLINGNESS TO PAY: A PARTIAL LEAST SQUARES . . . . .	519
<i>Karsten Lübke, Christian Hose and Thomas Obermeier</i>	
INSPECTING THE QUALITY OF ITALIAN WINE THROUGH CAUSAL REASONING . . . . .	521
<i>Eugenio Brentari, Maurizio Carpita and Silvia Golia</i>	
EXPLORING SOCIO-ECONOMIC FACTORS ASSOCIATED WITH ADHERENCE TO THE MEDITERRANEAN DIET: A MULTILEVEL APPROACH . . . . .	525
<i>Tiziana Laureti and Luca Secondi</i>	
BIG DATA AND 'SOCIAL' REPUTATION: A FINANCIAL EXAMPLE . . . . .	529
<i>Paola Cerchiello</i>	
BAYESIAN NETWORKS FOR STOCK PICKING . . . . .	533
<i>Alessandro Greppi, Maria Elena De Giuli and Claudia Tarantola</i>	
PORTFOLIO SELECTION WITH LASSO ALGORITHM . . . . .	537
<i>Riccardo Bramante, Silvia Facchinetti and Diego Zappa</i>	
SUNSPOT IN ECONOMIC MODELS WITH EXTERNALITIES . . . . .	540
<i>Beatrice Venturi and Alessandro Pirisinu</i>	

# DIAGNOSTIC TOOLS FOR GAMLSS FITTED OBJECTS

Andrea Marletta<sup>1</sup>, Mariangela Sciandra<sup>1</sup>

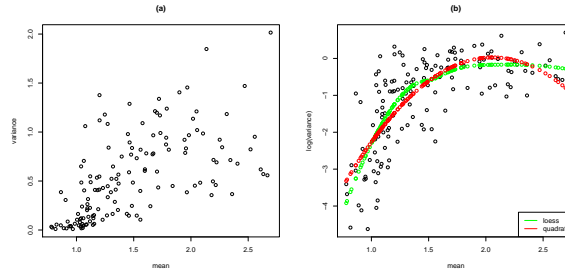
<sup>1</sup> Dipartimento di Scienze Economiche Aziendali e Statistiche, Università degli Studi di Palermo, (e-mail: [andrea.marletta@unipa.it](mailto:andrea.marletta@unipa.it), [mariangela.sciandra@unipa.it](mailto:mariangela.sciandra@unipa.it))

**ABSTRACT:** In the last years GAMLSS models were applied in many research fields representing a good solution to analyze data with huge variability. In this paper we propose a new approach to diagnostics in GAMLSS as an alternative to classical worm plot. An application will be shown where the class of GAMLSS is applied in order to detect the presence of liver fibrosis as a function of patients risk factors.

**KEYWORDS:** GAMLSS, liver fibrosis, mixture, worm plot, residuals analysis.

## 1 Introduction

We discuss some diagnostic tools for Generalized Additive Models for Location Scale and Shape (GAMLSS) in order to be able to identify possible departures from the model assumptions. Studying the adequacy of a GAMLSS model is not so obvious and little has been done in literature because of the several simultaneous assumptions each model includes about the different parameter involved in the model. So, for example, GAMLSS could show inaccuracies in the assumed linear predictors, one for each specified parameter or inadequacies related to overdispersion and misspecification in link functions. Moreover, due to the wide flexibility GAMLSS offer another common problem is related to misspecification of the family of conditional distribution. In this work we want to emphasize the problem of overdispersion, the most common form of unexpected variation. It occurs when the data exhibit variability exceeding that prescribed by the assumed distribution. As Fig. 1 shows, in the *liver fibrosis* example, data seem to be overdispersed. There is an increase of the variance with the mean ((a)) and a non linear relationship between means and the log transformed variances, as the two fitted curves show ((b)). In order to detect the presence of a variance-mean relationship, in Section 3 we propose the use of a mixture model using GAMLSS family of distributions.



**Figure 1.** Mean-variance (a) and mean-log(variance) (b) relationships in liver fibrosis data

## 2 The GAMLSS models

General Additive Models for Location Scale and Shape were introduced firstly by Rigby and Stasinopoulos (2001) as a way of overcoming some of the limitations associated with Generalized Linear Models and Generalized Additive Models. They represent a flexible class of models for several reasons. Firstly, the distribution of the response variable can be selected by a very wide range of distributions including highly skewed and kurtotic continuous and discrete distributions. Moreover, once the response distribution has been fixed, they allow to model all the parameters of the chosen distribution using parametric and/or non parametric smooth functions of the explanatory variables. So, assuming the response variable  $Y$  to follow a four parameters distribution  $Y \sim D(\theta)$  with  $\theta = (\mu, \sigma, \nu, \tau)$ , where  $\mu$  and  $\sigma$  usually are location and scale parameters while  $\nu$  and  $\tau$  shape parameters, the formulation of GAMLSS given by Rigby and Stasinopoulos (2005) is

$$g_k(\theta_k) = \eta_k = \mathbf{X}_k \beta_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}) \quad k = 1, 2, 3, 4 \quad (1)$$

where  $g_k(\cdot)$  are known monotonic link functions relating in a parametric way the distribution parameters to the explanatory variables  $\mathbf{X}_k$  and  $h_{jk}$  represent the non-parametric additive terms. The vector of parameters  $\beta_k$  and the non parametric terms can be estimated following several approaches as described in Rigby and Stasinopoulos (2005).

### 3 Diagnostic tools for GAMLSS models

As in a classical model regression framework, once a model is fitted, next step deals with the problem of model selection. In a GAMLSS setting, model selection is usually performed by comparing various competing models in which different combinations of the components change; then, the overall adequacy of the selected model is assessed through the analysis of the *randomized quantile residuals* (Dunn & Smyth, 1996). They are defined as

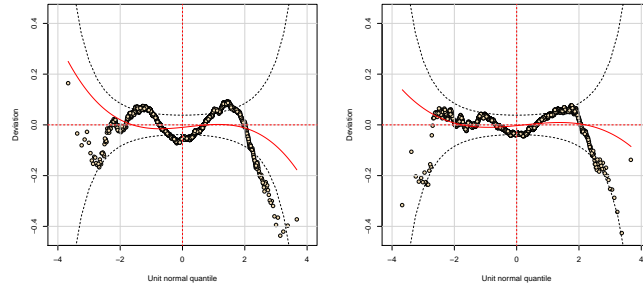
$$r_i^q = \Phi^{-1}(u_i), \quad i = 1, \dots, n$$

where  $\Phi(\cdot)$  represents the standard normal distribution function, and  $u_i$  is an uniform random variable on a specific region of the linear predictor. Strange pattern in the plot of these residuals against the predictors could suggest misspecified link functions. In order to identify regions of explanatory variables within which the models do not show an adequate fit an useful tool is given by the *worm plots* of residuals introduced by Van Buuren *et al.* (2001). The tool consists of a number of detrended Q-Q plots splitted according to some predictors. A model that fits the data well is characterized by “flat worms”.

This paper proposes the use of a mixture approach for GAMLSS when standard diagnostic tools show overdispersed data. The central idea is: if the worm plot shows for example M-shape pattern, it could suggest bimodality in statistical units. Then, GAMLSS mixture model could be used to identify the underlying distributions.

### 4 A real dataset example

Liver fibrosis is one of the ten most frequent causes of death in the world and consists in a massive presence of connective tissue. It can be classified in 5 stages through the Metavir scoring system from a normal ( $F0$ ) to a cirrhotic ( $F4$ ) liver. In medicine, liver biopsy represents the gold standard test for staging liver disease. An alternative diagnostic technique is represented by the Acoustic Radiation Force Impulse (ARFI). ARFI measures the liver stiffness through mechanical excitation of tissue using acoustic pulses producing shear waves propagation. The ARFI principle is that the stiffer the tissue, the faster be shear waves propagate. The dataset used in this example contains data about ARFI measurements collected in 2013 for 141 patients. To each elastography are associated a different number of measurements so the dataset shows a two-step hierarchical structure: a level for the exam and a second level for the measurements done during the same exam. The response variable is liver stiffness



**Figure 2.** Worm plots for two fitted GAMLSS objects

(measured as wave speed in m/s) while explanatory variables are divided into two groups: patient specific explanatory variables (sex, age, size and weight) and explanatory variable of exam (depth, liver segment, patient position). In figure 2 two worm plots are shown: on the left a simple GAMLSS with a four parameters BCPE distribution shows a clear M-shape pattern; on the right we use a mixture to obtain a more flat worm and more points between boundaries. A possible hypothesis is that two groups could represent two classes of subjects: healthy and cirrhotic patients. To find further evidence about these conclusions, we will try to simulate data from different scenarios, for example by considering several degrees of overdispersion. Moreover, simulations will be also used to study the inferential properties of the proposed approach.

## References

- DUNN, P.K., & SMYTH, G.K. 1996. Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, 236-244.
- RIGBY, R.A., & STASINOPOULOS, D. M. 2005. Generalized additive models for location, scale and shape). *Applied Statistics*, 54(3), 507-544.
- RIGBY, R.A., & STASINOPOULOS, D. M. 2009. A flexible regression approach using GAMLSS in R.
- VAN BUUREN, S., & FREDRIKS. 2001. Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in medicine*, 1259-1277.