



Classification and Data Analysis 2007

Book of Short Papers

Meeting of the Classification and Data Analysis
Group of the Italian Statistical Society



eum > economia > statistica

eum > economia > statistica

The Sixth Scientific Meeting of the CLAssification and Data Analysis Group (CLADAG) of the Italian Statistical Society was held in Macerata, September 12th-14th, 2007.

Conference web site: <http://cladag2007.unimc.it>

The present book contains the short papers presented during this meeting.

eum edizioni università di macerata



ISBN 978-88-6056-020-9



9 788860 560209

€ 40,00

CLASSIFICATION AND DATA ANALYSIS 2007

Book of Short Papers
Meeting of the
CLASSIFICATION and Data Analysis Group
of the Italian Statistical Society



MACERATA - SEPTEMBER, 12TH-14TH, 2007

eum

Isbn 978-88-6056-020-9
Prima edizione: settembre 2007
© 2007 eum edizioni università di macerata
Vicolo Tornabuoni, 58 - 62100 Macerata
info.ceum@unimc.it
<http://ceum.unimc.it>
Realizzazione e distribuzione:
Quodlibet società cooperativa
Via S. Maria della Porta, 43 - 62100 Macerata
www.quodlibet.it
Stampa: Grafica Edirice Romana s.r.l., Roma

This book was realized with the financial support of



Università degli Studi
di Macerata



REGIONE
MARCHE



SISTEMA
INFORMATIVO
STATISTICO

Table of contents

<i>Preface</i>	19
----------------	----

KEY-NOTE PAPERS

S. R. Masera, G. Mazzoni <i>Actuarial and Continuous Time Risk Models: Towards a Synthesis, Changing financial paradigms in the new enterprise economy</i>	23
G. McLachlan <i>Clustering of High-Dimensional and Correlated Data</i>	29
A. Montanari <i>Classification by mixture and latent variable models</i>	37
A. Rizzi <i>Statistical Methods for Cryptography</i>	43
R. Zang, H. Bozdogan <i>Model Selection in Relevance Vector Machines (RVMs) Using Information Complexity and Genetic Algorithm (GA)</i>	51

INVITED PAPERS

Specialized Session 1

Knowledge extraction from temporal data models

Organizer: D. Piccolo

R. Baragona, S. Vitrano <i>Statistical and numerical algorithms for time series classification</i>	65
M. Corduas <i>Comparing time series: shape-based or structural similarities?</i>	69
G. Scepi, G. Milone <i>Temporal Data Mining: clustering methods and algorithms</i>	73

Specialized Session 2

Statistical models with errors-in-covariates

Organizer: A. Pastore

- M. Battauz, R. Bellio
Structural analysis of linear mixed models with measurement error 79
- A. Guolo
Measurement error correction techniques in case-control studies 83
- P. Mantovan, A. Pastore
Dynamic regression with covariate errors by covariate local clusters 87

Specialized Session 3

Multivariate Analysis for Microarray Data

Organizer: M Vichi

- M. Alfö, F. Martella, M. Vichi
Biclustering of microarray data 93
- A. M. Mineo, L. Augugliaro, C. Fede, M. Ruggieri
A Statistical Calibration Method Based on Non-Linear Mixed Model for Affymetrix Probe Level Data 97
- E. Wit, G. Green
Random effects modelling for multivariate data from cDNA microarrays 101

Specialized Session 4

Cluster Analysis of complex data

Organizer: R. Verde

- A. Irpino, R. Verde
Clustering linear models using Wasserstein distance 107
- F. Palumbo, A. Iodice D'Enza
A two-step iterative procedure for clustering of binary sequences 111
- A. Petrucci, C. T. Brownless
Spatial Clustering Methods for the Detection of Homogenous Areas 115

Specialized Session 5

Educational Processes Assesment by means of Latent Variables Models

Organizer: S. Mignani

- M. Battauz , E. Gori
Educational assessment in presence of heteroscedastic measurement errors 121

B. Chiandotto, F. Polverini, B. Bertaccini <i>The effectiveness of university education: a structural equation models</i>	125
W. J. van Der Linden <i>Response Times on Test Items: Models and Applications</i>	129

Specialized Session 6

Classification of complex data

Organizer: D.A. Zighed

H. Azzag, C. Guinot, G. Venturini <i>On data clustering with bio-inspired algorithms</i>	135
F. de Carvalho, Y. Lechevallier, R. Verde <i>Clustering approach on interval data</i>	139
B. Fichet <i>Around the ultrametric sandwich</i>	143
D. A. Zighed <i>Separability of classes in a multidimensional space</i>	147

Specialized Session 7

Multidimensional Scaling

Organizer: A. Okada

G. Bove <i>Models for asymmetry in proximity data</i>	153
T. Imaizumi <i>On extracting a local structure of asymmetric matrix</i>	157
A. Okada <i>Two-Dimensional Centrality of Asymmetric Social Network</i>	161

Specialized Session 8

Statistical Models for Public Policies

Organizer: S. Ingrassia

M. Caserta <i>Enlarging internal regional markets: measuring the effects on local development</i>	167
R. Innocenti, A. Giommi, C. Brownless <i>A New Statistical Zoning for the Municipality of Firenze</i>	171
M. Riani, R. Lagomarsini, A. Micozzi <i>Robust Clustering for Performance Evaluation</i>	175

Specialized Session 9

Classification models for enterprise risk management

Organizer: P. Giudici

- E. Bonafede, P. Cerchiello
A proposal to fuzzify categorical variables in operational risk management 185
- C. Cornalba
Statistical models to measure IT operational risks 189
- D. Fantazzini, S. Figini
Predictive dynamic models for SMEs 193

Specialized Session 10

Model based clustering

Organizer: D. Vicari

- C. Biernacki, A. Lourme
*Simultaneous Model-Based Clustering of Data Arising from Different
Populations* 199
- G. Galimberti, G. Soffritti
*Multiple cluster structures and mixture models: recent developments for
multilevel data* 203
- D. Vicari, M. Vichi
Multivariate regression model with clustering of objects and variables 207

CONTRIBUTED PAPERS

Solicited Papers Session 1

Applications in Classification and Segmentation

Organizer: J. Antoch

- J. Antoch
Testing the Difference of the ROC Curves in Bigamma Model 217
- C. Conversano, E. Dusseldorp
Classification Trunk Approach for Variable Selection and Threshold Interactions 219

J. Klaschka <i>Tree-based classification of EEG spectra</i>	223
R. Miele <i>An Ant Colony-based Algorithm for Classification and Regression Trees"</i>	227
R. Siciliano, V. A. Tutore, M. Aria <i>3Way Trees</i>	231

Solicited Papers Session 2

Classification Issues in Social Network Analysis

Organizer: G. Giordano, M. P. Vitale

V. Batagelj <i>Clustering in Networks</i>	237
G. Giordano, M. P. Vitale <i>Local Factorial Analysis and Contiguity Constraints in Social Networks</i>	241
Y. H. Said, E. J. Wegman, W. Sharabati, J. T. Rigsby <i>Social Networks of Author-Coauthor Relationships</i>	245
S. Zaccarin, G. Rivellini <i>Modelling network data: an Introduction to Exponential Random Graph Models</i>	249

Solicited Papers Session 3

Metainformation and knowledge extraction from textual data bases

Organizer: S. Balbi

S. Bolasco, P. Pavone <i>Automatic dictionary and rule-based systems for extracting information from text</i>	255
A. Canzonetti <i>Semantic classification and cooccurrences: a method for the rules production for the information extraction from textual data</i>	259
F. della Ratta, B. Lorè, G. La Rocca <i>Textual analysis perspectives on categorisation of activities in Istat survey on occupations</i>	263
G. Infante, M. Misuraca <i>Text Mining Strategies for Analyzing Semi-structured Corpora</i>	267

Solicited Papers Session 4
Customer Relationship Management
Organizer: C. Davino

F. Camillo, C. Liberati	
<i>A micro-data mining approach for qualitative-emotional marketing using neuro-information</i>	273
C. Davino, R. Del Gobbo	
<i>Structural Neural Networks for Modeling Customer Satisfaction</i>	277
S. Figini, A. Roccato	
<i>How to improve predictive models for database marketing applications</i>	281

Solicited Papers Session 5
Missing Data
Organizer: A. Plaia

M. Aria, A. D'Ambrosio, R. Siciliano	
<i>Robust Incremental Trees for Missing Data Imputation and Data Fusion</i>	287
A. Plaia, A. L. Bondi	
<i>Regression imputation for Space-Time datasets with missing values</i>	291
I. Sulis, M. Porcu	
<i>A multiple imputation approach in a survey on university teaching evaluation</i>	295

Solicited Papers Session 6
Statistical methods in decision-making
Organizer: R. Amenta

G. Adelfio, M. Chiodi, D. Luzio	
<i>An algorithm for earthquakes clustering based on maximum likelihood</i>	301
F. Di Salvo, N. Ferotti, S. Consagra	
<i>Hospital performance comparison: assessing inappropriate stay in the hospitals of Palermo</i>	305
C. Gaetan, L. Greco	
<i>Weighted Likelihood Inference in Gaussian Spatial Linear Models</i>	309

Solicited Papers Session 7
Multivariate methods in social research
Organizer: D. F. Iezzi

M. Civardi, F. Crippa	
<i>From University to Work: a measure of coherence between education and job</i>	315

P. Costantini		
<i>Analyzing learning effects through Latent Growth Models</i>		319
L. Fabbris		
<i>Dimensionality of scores obtained with a pair-comparison tournament system of questionnaire items</i>		323
D. F. Iezzi, M. Grisoli		
<i>Using Rasch measurement to assess the role of the traditional family in Italy</i>		327

Solicited Papers Session 8

Statistical Models with Latent Features

Organizer: L. Tardella

S. Arima, L. Tardella		
<i>A Bayesian Approach to Peabody Picture Vocabulary Test Revised</i>		333
F. Bartolucci, A. Farcomeni		
<i>Dynamic logit models for panel data based on a latent Markov heterogeneity structure</i>		337
G. Petris, S. Petrone		
<i>Bayesian inference for dynamic linear models with random variances</i>		341
R. Rocci, A. Maruotti		
<i>An INDSCAL based mixture model</i>		345

Solicited Papers Session 9

Classification Trees

Organizer: R. Siciliano, M. Aria

C. Conversano, F. Mola		
<i>Sequential Automatic Search of a Subset of Classifiers for Multi-Attribute Response Prediction</i>		351
G. Galimberti, M. Pillati, G. Soffritti		
<i>Comparing strategies for robust regression tree construction</i>		355
V. A. Tutore, A. Mooijaart		
<i>Optimal Scaling Trees</i>		359
M. Vezzoli, C. Stone		
<i>CRAGGING</i>		363

Solicited Papers Session 10
Exploring data structures with the Forward Search
Organizer: A. Cerioli

M. Bini, B. Bertaccini <i>Robust fuzzy classification</i>	369
D. G. Calò <i>Outlier detection via Forward Search: a proposal based on mixture models</i>	373
D. Perrotta, F. Torti <i>Detecting price outliers in European trade data with backward and forward methods</i>	377
N. Solaro, M. Pagani <i>The Forward Search for Metric Multidimensional Scaling</i>	381

Contributed Papers Session 1
Clustering in Marketing Research and Data Mining

R. Menichelli, F. Palumbo <i>Clustering of binary data in AR mining: a comparison among different strategies</i>	387
I. Morlini, S. Zani <i>Comparing approaches for clustering mixed mode data: an application in marketing research</i>	391
G. Schoier, B. Bato <i>A modification of the DBSCAN Algorithm in a Spatial Data Mining Approach</i>	395
A. Tarsitano, I. Bonafine <i>Distance function for mixed type data</i>	399
A. F. X. Wilhelm <i>Association rule mining of multimedia content</i>	403

Contributed Papers Session 2
Data Analysis in Finance

A. Amendola, M. Niglio, C. Vitale <i>Temporal aggregation and closure of VARMA models. Some new results</i>	409
L. Attardi, D. Vistocco <i>Evaluating financial portfolio activeness through style analysis models</i>	413

R. Castellano, L. Scaccia <i>Bayesian hidden Markov models for financial data</i>	417
S. Cieply, R. Abdesselam <i>Interrelations between Asset and Financial Structures: the case of French NTIC firms</i>	421
F. Domma, S. Giordano, P. F. Perri <i>Modelling the dependence structure of financial assets using copula functions</i>	425
M. Marozzi, L. Santamaria <i>How to Make a Complex Corporate Finance Indicator Simpler</i>	429

Contributed Papers Session 3

Mixture and Latent Class model

M. Alfò, A. Maruotti <i>A Hierarchical Model for Time Dependent Longitudinal Data</i>	435
P. Coretto, C. Hennig <i>Choice of the improper density in robust improper ML for finite normal mixture</i>	439
G. E. Montanari, M. G. Ranalli, P. Eusebi <i>Multilevel Latent Class Models for evaluation of long term care facilities</i>	443
R. Piccarreta, M. Bonetti, S. Venturini <i>Clustering-based measurement of dependence</i>	447
L. Scrucca <i>Visualization of model-based clustering structures</i>	451
K. Yamaguchi, M. Watanabe <i>Mixture models for multivariate count data</i>	455

Contributed Papers Session 4

Economics

V. Adorno, C. Bernini, G. Pellegrini <i>Comparing matching methods in policy evaluation</i>	461
C. Cappelli, F. Di Iorio <i>Structural breaks versus long memory, a simulation study</i>	465
D. De Vos, K. Vanhoof, H. Van Landeghem <i>Symbolic Data As Measures For Logistical Performances</i>	469
C. Giusti <i>Multiple imputation of income for the Labour Force Survey</i>	473

Contributed Papers Session 5

Latent Variables

- G. Adelfio
Earthquakes clustering interpretation based on a second-order diagnostic approach 479
- S. Bianconcini, S. Cagnone
A Multilevel Latent Variable Model for Multidimensional Longitudinal Data 483
- E. Otranto
A time varying Hidden Markov Model with latent information 487

Contributed Papers Session 6

Non Parametric Regression

- D. Coin
A test for Normality with high Power against Symmetrical Alternatives 493
- L. De Benedictis, M. Callegati, M. Tamperi
Semiparametric analysis of the specialization-income relationship 497
- R. Miglio
Non Parametric Procedures to Explore Interactions in Epidemiological Studies 501
- A. Zirilli, A. Alibrandi
SMART Regression: an application to multidimensional data in a medical context 505

Contributed Papers Session 7

Cluster Structures

- A. Alibrandi, M. Giacalone
A Comparison among Statistical Criteria to Establish the Number of Clusters in Multidimensional Data 511
- A. Balzanella, E. Romano, R. Verde
Outliers detection for a Robust Curves Clustering Algorithm 515
- A. Iodice D'Enza
Visual Monitoring Tool of Association Patterns in Binary Data Flows 519
- G. Notarstefano, R. Scuderi
Polarization and Dynamic Clustering: A Review 523
- S. Poletini, L. Di Consiglio
Disclosure risk estimation with survey data: a comparative study 527

Contributed Papers Session 8

Customer Satisfaction

L. Berzieri <i>Demographic snapshot of Parma electoral colleges</i>	533
P. Chirico <i>Common Optimal Scaling for Customer Satisfaction Multidimensional Models</i>	537
M. Civardi, E. Zavarrone <i>Ordinal Scales and Customer Satisfaction: an Applet</i>	541
P. Ferrari, C. V. Fiorio, L. Pagani <i>A two-step procedure to analyze users' satisfaction</i>	545

Contributed Papers Session 9

Discrimination

R. Abdesselam <i>Mixed Discriminant Analysis</i>	551
E. Bricchieri Colombi, C. T. Brownlees, G. M. Gallo <i>Predicting Binary Outcomes Using Flexible and Parsee's Methods</i>	555
S. Ingrassia, I. Morlini <i>Computational experiences with the equivalent number of degrees of freedom of neural networks</i>	559
R. Sanzo, E. Trinca <i>A multivariate approach to non-response treatment for Foreign control companies Survey</i>	563

Contributed Papers Session 10

Robust Classification

A. Corbellini, P. Ganugi, L. Crosato <i>Robust sequential Fitting of the Pareto II distribution: Theoretical and Computational Aspects</i>	569
M. Daniele, A. Bondi <i>Outliers in hierarchical models: an application to air pollution data</i>	573
A. Gottard, S. Pacillio <i>Robustifying SINful procedure</i>	577
P. Mascia, R. Miele, F. Mola <i>Robustness Issues in Classification and Regression Trees</i>	581
S. Scippacercola <i>The Progressive Single Linkage Algorithm Based on Minkowski Ultrametrics</i>	585

Contributed Papers Session 11

Bayesian Network

- A. Brogini, D. Slanzi
Several computational studies about variable selection for Bayesian networks 591
- M. Mascherini
Exploring eBusiness Readiness Indicators using Bayesian Networks 595
- L. Scaccia, F. Bartolucci
Bayesian inference for marginal models under equality and inequality constraints 599

Contributed Papers Session 12

Textual data

- M. Dimai, N. Torelli
Clustering textual data by Latent Dirichlet Allocation: Applications and Extensions to Hierarchical Data 605
- D. F. Iezzi, M. Giorgino
Intimate femicide in Italy: a model to classify how killings happened 609
- S. Volo, C. Fisichella
Evaluating tourists experiences 613

Contributed Papers Session 13

Data Analysis

- M. Greenacre
Power Transformations in Correspondence Analysis and Related Methods 619
- G. Menardi, N. Torelli
Multimodal projection pursuit using the adjusted critical bandwidth 623
- R. Romano, T. Naes, P. Brockhoff
The use of analysis of variance and three-way factor analysis methods for studying the quality of a sensory panel 627

POSTER SESSION

- R. Belicchi, F. Crippa, L. Cesana, R. Daini
Lezak TinkerToy Test: freely making constructions to assess executive functions. An Italian normative sample 633

A. Belliggiano, S. Staffieri <i>Private label organic food products: profile and behaviour of Italian consumers characterized through multivariate approach</i>	637
G. Boari, G. Cantaluppi, A. De Lauri <i>Handling missing values in PLS path modelling: comparison of alternative procedures</i>	641
S. Bozza, F. Taroni, F. Guy, M. Schmittbuhl <i>Taxonomic membership in great apes: a probabilistic approach from mandibular morphology</i>	645
S. De Cantis, M. Ferrante <i>Data quality and errors correction in Italian official statistics on guests of accommodation establishments</i>	649
C. Fukae <i>The Myths of Global Warming and Nuclear Power: Evidence from public opinion survey in Japan</i>	653
R. Gismondi, A. R. Giorgi, T. Pichiorri <i>Reducing Sampling Error using Deterministic Sample Selection in the Italian Retail Trade Monthly Survey</i>	657
R. Gismondi, M. A. Russo <i>Selection and Statistical Treatment of Variables for Evaluating Quality of Life in the Italian Provinces</i>	661
U. Magagnoli, P. M. Chiodini <i>Unilateral Hypothesis Tests of Efficiency Functions with Heteroscedastic Errors: an Iterative Procedure</i>	665
A. M. Olivieri <i>Excursionist flows and the attractiveness on the Aeolian Islands</i>	669
A. G. Quaranta, S. Fedeli, E. Voltattorni <i>Basel 2: firm's Scoring via Cluster Analysis and Artificial Neural Networks</i>	673

Ph.D. SPECIAL SESSION

G. Adelfio <i>Point processes residual analysis and asymptotical distribution of transformed versions of some second-order statistics</i>	679
V. Adorno <i>Program evaluation with continuous treatment: theoretical considerations and empirical application</i>	683

<i>Health related Quality of Life: some methodological aspects of Rasch analysis</i>	689
P. Eusebi <i>Many Facets Rasch Models: a Hierarchical Latent Class approach</i>	693
M. Matteucci <i>A multidimensional Item Response Theory approach for the University guidance</i>	697
R. Ragona <i>A State-Space Model for the Generation of Spatial Maps</i>	701
E. Romano <i>A new strategy for curves clustering based on free knots spline estimation</i>	705
I. Suliis <i>Measuring students' assessment on 'university course quality' using mixed-effects models</i>	709
Indice degli autori	713

Data quality and errors correction in Italian official statistics on guests of accommodation establishments ⁽¹⁾

Stefano De Cantis, Mauro Ferrante

Dipartimento di Metodi Quantitativi per le Scienze Umane

Università degli Studi di Palermo

Viale delle Scienze – Edificio 13, 90133 Palermo

decantis@unipa.it, mauroferrante@unipa.it

Abstract: This paper discusses the quality of data in Italian survey on guests of accommodation establishments by looking at some basic conditions of reliability of the elementary records. Monthly data referred to the 32 Sicilian tourism promotion agencies for seven years are checked and the main kinds of error are presented. Some hypothesis with reference to the sources of the errors and to their relative correction procedures are discussed. The method proposed, relatively simple to apply, could be implemented as a part of the standard data pre-processing.

Keywords: *Tourism statistics, error rate, reliability*

1. Introduction

One of the most important official Italian survey on tourism is the so-called *Survey on guests of accommodation establishments* (ISTAT, 2004). It is usual to refer to these data as information on tourists' arrivals (ISTAT, 2006, p.209). This is a conceptual mistake for several aspects. First of all because information about characteristics of guests are not available, this makes impossible to distinguish tourists from other guests (European Commission, 2007, *Technical notes*). Secondly, there is a huge number of accommodation establishments (such as private accommodations), not registered to the Italian Commercial Facilities Register for which information on guests are not covered by this survey. Moreover, the official accommodation establishments cannot declare the real number of arrivals and/or of presences, to avoid direct or indirect taxation. Finally, and even more important, a single guest determines more than one arrival on a given destination for a given month, if he/she changes accommodation establishment during his/her trip, generating errors resulting in an overestimate of the number of tourists' arrivals, and consequently in an underestimate of the average length of stay (Vaccina and Parroco, 2006). This problem is well documented but, unfortunately, there is no direct way to improve the reliability of the survey on measuring the number of tourists without changing the framework of the survey (for some pioneers researches on a single local district, see Parroco and Vaccina, 2006). In this paper we investigate the way to improve the quality of data controlling for several basic conditions, by checking the internal coherence of records.

⁽¹⁾ The present paper is a common work and responsibility of both authors, however S. De Cantis wrote sections 1 and 3, and M. Ferrante wrote sections 2 and 4.

2. Detecting errors

The arrivals (number of guests spending at least one night) and the presences (number of nights spent by guests) in accommodation establishments are recorded by Italian tourism promotion agencies, according to the so-called Istat CTT/1 model (ISTAT, 2004). It is constituted by four main sections derived from the combination of two factors: region of origin of guests, and typology of accommodation used. According to our framework of analysis, the elementary record is composed by a pair of values as follows:

$$(x_{t,w,i,j}, y_{t,w,i,j}) \text{ with } x_{t,w,i,j} \text{ and } y_{t,w,i,j} \in \mathbb{N}^+ \quad \forall t, w, i, j \quad (1)$$

where: x and y are the arrivals and the presences on accommodation establishments; t (time) ranges from 1 (jan-1999) to 84 (dec-2005); w represents the 32 Sicilian tourism promotion agencies; i indicates the country (if foreign) or the region (if Italian) of origin of guests, and it ranges from 1 to 76 (21 Italian regions and 55 foreign countries); and j represents the ten typologies of accommodation establishments, resulting in a total number of pairs equals to 2,042,880. Since one arrival generates at least one presence, it holds:

$$x_{t,w,i,j} \leq y_{t,w,i,j} \quad \forall t, w, i, j \quad (2)$$

If we define the average length of stay (z) as the ratio between presences and arrivals, the condition (2) becomes:

$$z_{t,w,i,j} = y_{t,w,i,j} / x_{t,w,i,j} \geq 1 \quad \forall t, w, i, j \quad (3)$$

To identify all the possible violations of the condition (3), monthly pairs of arrivals and presences referred to Sicily from 1999 to 2005 were classified according to eight possible conditions, as shown in Table 1. If we exclude records referred to condition A (null pairs), only the records corresponding to the conditions B and C can be considered as correct. On the other side, records falling into conditions from D to H must be considered as incorrect records. As result, we found a raw error rate equals to 2.4%. It should be noted that we also found several records falling into condition B, that are "pathological". For example, 1,117 (0.3%) records have an average length of stay greater than 60 days.

Table 1: Number of correct and incorrect records according to different conditions on values of arrivals and presences in Sicily (1999-2005).

	Conditions on			Number of pairs	%	%
	Arrivals	Presences	Average length of stay			
A	$x_{t,w,i,j} = 0$	$y_{t,w,i,j} = 0$	Undefined	1,602,035		78.4%
B	$x_{t,w,i,j} > 0$	$y_{t,w,i,j} > 0$	$y_{t,w,i,j} / x_{t,w,i,j} > 1$	347,985	78.9%	
C	$x_{t,w,i,j} > 0$	$y_{t,w,i,j} > 0$	$y_{t,w,i,j} / x_{t,w,i,j} = 1$	82,293	18.8%	
D	$x_{t,w,i,j} = 0$	$y_{t,w,i,j} > 0$	$y_{t,w,i,j} / x_{t,w,i,j} \rightarrow +\infty$	9,764	2.2%	
E	$x_{t,w,i,j} > 0$	$y_{t,w,i,j} = 0$	$y_{t,w,i,j} / x_{t,w,i,j} = 0$	252	0.1%	
F	$x_{t,w,i,j} > 0$	$y_{t,w,i,j} > 0$	$y_{t,w,i,j} / x_{t,w,i,j} < 1$	548	0.1%	
G	$x_{t,w,i,j} > 0$	$y_{t,w,i,j} < 0$	$ y_{t,w,i,j} / x_{t,w,i,j} > 1$	3	0.0%	
H	Other conditions			0	0.0%	
Subtotal(Conditions from B to G)				440,845	100.0%	21.6%
Total				2,042,880		100.0%

Source: Osservatorio turistico Regione Siciliana, and authors' calculations.

To include all the possible cases, it is possible to explicit the residual conditions H, as follows:

- H1: $x_{t,w,i,j} > 0$ and $y_{t,w,i,j} < 0$ and $|y_{t,w,i,j} / x_{t,w,i,j}| < 1$; H2: $x_{t,w,i,j} > 0$ and $y_{t,w,i,j} < 0$ and $|y_{t,w,i,j} / x_{t,w,i,j}| = 1$;
H3: $x_{t,w,i,j} < 0$ and $y_{t,w,i,j} < 0$ and $|y_{t,w,i,j} / x_{t,w,i,j}| > 1$; H4: $x_{t,w,i,j} < 0$ and $y_{t,w,i,j} < 0$ and $|y_{t,w,i,j} / x_{t,w,i,j}| = 1$;
H5: $x_{t,w,i,j} < 0$ and $y_{t,w,i,j} < 0$ and $|y_{t,w,i,j} / x_{t,w,i,j}| < 1$; H6: $x_{t,w,i,j} = 0$ and $y_{t,w,i,j} < 0$ and $|y_{t,w,i,j} / x_{t,w,i,j}| \rightarrow +\infty$
H7: $x_{t,w,i,j} < 0$ and $y_{t,w,i,j} > 0$ and $|y_{t,w,i,j} / x_{t,w,i,j}| < 1$; H8: $x_{t,w,i,j} < 0$ and $y_{t,w,i,j} > 0$ and $|y_{t,w,i,j} / x_{t,w,i,j}| = 1$;
H9: $x_{t,w,i,j} < 0$ and $y_{t,w,i,j} > 0$ and $|y_{t,w,i,j} / x_{t,w,i,j}| > 1$.

3. Describing the sources of error variability

The above classification is useful to determine the possible sources of error, and subsequently the more appropriate ways to correct the data, as it will be discussed into the last section. Since the distribution of errors on records resulted to be not uniform according to several dimensions of our data, in Table 2 we reported the raw error rates specific by tourism district agency, by months, by years, by typology of accommodation establishment, respectively. Moreover, to consider the joint effect of these variables we estimated, only for descriptive purposes, several logit models. Let $n_{1,m,v,w,i,j}$ and $n_{2,m,v,w,i,j}$ the number of incorrect and correct records respectively, for the m -th month, for the v -th year, for the w -th tourism district, for the i -th nationality (1 = Italian, 2 = Foreign), and for the j -th accommodation establishment. The logit model for the probability of incorrect records – with a simple structure, and the best goodness of fit (interactions between factors did not appeared interesting) – can be expressed as follows:

$$\log\left(\frac{n_{1,m,v,w,i,j}}{n_{2,m,v,w,i,j}}\right) = \text{logit}\left(n_{1,m,v,w,i,j}\right) = \theta + \theta_m^M + \theta_v^V + \theta_w^W + \theta_i^I + \theta_j^J \quad (4)$$

$\forall m = 1, 2, \dots, 12; v = 1, 2, \dots, 7; w = 1, 2, \dots, 32; i = 1, 2; j = 1, 2, \dots, 10$ and $\theta_{12}^M = \theta_{12}^V = \theta_{12}^W = \theta_{12}^I = \theta_{12}^J = 0$.

Through the model (4) the effect of each factor is evaluated controlling for the other ones, and Table 2 reports the corresponding parameters (Odds Ratios, OR). However, the results are almost comparable. We found a different raw error rate between records related to Italian guests and foreign ones (1.9% vs 2.8%, respectively), and a risk of incorrect record almost double in the case of Italian guests, compared to the foreign ones. A large variation appears in errors rate and in odds ratios, with reference to different tourism districts. Agrigento resulted to be the district with the lowest associated risk (adj. OR = 0.02, rank 32), considering as reference (OR = 1) the district of Trapani (others municipalities).

Table 2: Raw error rates and adjusted odds ratios for incorrect records.

Destination	Agrigento	Agrigento others	Sciaccò	Gela	Calimussina	Castelluzzo (altres)	Aciroale	Collegreone	Catania - Scacalo	Nicolosi	Catania (altres)	Enna	Piazza Armerina	Enna (Others)	Capo d'Orlando	Giardini	Strozz	Isole Eolie	Mazara	Milazzo	Patti	Taormina	Mazara (altres)	Cybalta	Taormina (altres)	Mazara (altres)	Ragusa	Ragusa (altres)	Siracusa	Siracusa (altres)	Ereje	Trapani	Trapani (altres)
Error rate	0.02	1.76	3.21	4.25	0.59	2.32	0.07	0.88	1.50	1.20	4.58	1.03	0.94	5.07	2.79	2.46	3.37	6.77	1.55	2.11	2.13	2.37	2.89	2.68	2.40	0.10	0.16	3.85	3.13	2.23	1.53	1.89	
Adj. OR	0.02	1.14	3.14	4.44	0.57	2.07	0.05	0.64	0.67	0.48	2.94	0.99	0.57	4.29	2.25	2.13	1.89	4.17	1.24	1.61	1.17	1.65	1.58	1.24	1.48	0.05	0.07	1.89	1.84	1.78	1.34	1.00	
Rank	32	21	4	1	26	8	30	25	24	28	5	22	26	2	6	7	9	3	18	14	20	13	15	18	16	30	29	9	11	12	17	22	

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Error rate	4.77	2.59	2.16	1.67	1.73	1.89	1.82	2.26	2.25	2.90	3.48	2.72
Adj. OR	1.93	0.99	0.76	0.53	0.54	0.57	0.52	0.66	0.67	0.94	1.31	1.00
Rank	1	4	6	11	10	9	12	8	7	5	2	3

Year	1999	2000	2001	2002	2003	2004	2005
Error rate	2.87	2.62	2.46	2.43	2.01	2.15	2.40
Adj. OR	1.46	1.35	1.23	1.17	0.90	0.92	1.00
Rank	1	2	3	4	7	6	5

Accommodation establishment typology	Hotels						Rented facilities	Camping & resorts	Rural tourism facilities	Other establishments
	5 stars	4 stars	3 stars	2 stars	1 star	Residences				
Error rate	1.66	1.28	1.37	1.67	1.96	10.15	3.59	3.15	2.25	4.91
Adj. OR	1.63	0.98	1.00	1.20	1.40	11.53	3.48	3.18	2.08	4.63
Rank	6	10	9	8	7	1	3	4	5	2

Error rates demonstrate also a dependence on months, with relatively lower values in high season months (e.g. Apr.-Sep.), and with higher values in low season months (e.g. Nov.-

Feb.). Instead, during the years considered, it seems not to be strong differences, both in the error rates and in the odds ratios. Finally, considering different accommodation establishments, it is possible to highlight a strong difference between the relatively small raw rates of hotels from 5 to 1 star, and the other typologies; with a maximum in correspondence of *Residences* (Error rate = 10,15%, OR = 11.53).

4. Considerations on errors correction and conclusions

Recalling the conditions expressed in Table 1, it is possible to discuss some considerations related to the causes of errors and to their relative correction procedure. Conditions D and E are referred to the situation in which one of the two elementary data (arrivals or presences) is zero. In these cases, considering the zero value as a missing one, a procedure of imputation can be based on an "appropriate value" of the average length of stay as an indicator of the relationship between the existing value and the missing one. Condition F is not easy to solve in term of imputation procedure. We assume that these errors derive from a mixture of correct and incorrect (deriving, for example, from condition D) records. In this case if we suppose that the most reliable data is the one related to the arrivals, we will use, also in this case, an "appropriate value" of the average length of stay, to derive the correspondent value of presences. Finally, condition G can reveal a simple mistake on editing; thus a procedure of correction can consists in a simple substitution of the value of presences with its absolute value. With reference to the remaining conditions included in H, we could suggest the same procedure used for G, with reference to the sub-cases from H2 to H4, and H8-H9, while with reference to the other possible situation, here not discussed in detail, a combination of the procedures used for D, E, and F, could be adapted.

In conclusion, this paper showed how to control for reliability of the elementary records of Italian survey on guests of accommodation establishments, by checking jointly data on arrivals and presences. The results showed an absence of accurate controls on the data sent by the accommodation establishment' responsible to the local tourism promotion agency. The method proposed, relatively simple to apply, could be implemented as a part of the standard data pre-processing. This control could be improved through the use of other constraints on the supply side derived, for example, from the total number of bed places available. It is important to perform these controls before aggregating the data, since the process of aggregation could determine a subsequent impossibility of detecting errors. Moreover, since the presence of errors seems to have some systematic components, a deeper analysis could suggests the more appropriate way to eliminate the causes of errors or to impute appropriate values, improving the quality of data.

References

- European Commission (2007) *Tourism statistics*, Luxembourg, <http://ec.europa.eu/eurostat>
- ISTAT (2004) *Statistiche del turismo: Anno 2002*. Roma, www.istat.it
- ISTAT (2006) *Italian Statistical Abstract, 2004*. Roma, www.istat.it
- Parroco A.M., Vaccina F. (2006) Methodological issues and first results of a study to identify and evaluate the error of some tourism statistics. *Proceedings of XLIII Scientific Reunion of Italian Statistical Society (SIS)*, Torino. Cleup, Padova.
- Vaccina F., Parroco A.M. (2006) Referring to Space and Time when using territorial data: The case of Touristic Arrivals. *Proceedings of International Statistical Institute (ISI) 55th Session*, Sidney, Australia.