CLADAG 2013

9 th SCIENTIFIC MEETING OF THE CLASSIFICATION AND DATA ANALYSIS GROUP OF THE ITALIAN STATISTICAL SOCIETY

September 18 - 20, 2013
University of Modena and Reggio Emilia
San Geminiano Complex - Modena, Italy

Book of Abstracts

Editors: Tommaso Minerva, Isabella Morlini, Francesco Palumbo
CLEUP
ISBN: 9788867871179

Patronage

Dipartimento di Economia Marco Biagi
Dipartimento di Comunicazione ed Economia
Dipartimento di Scienze Fisiche, Informatiche e Matematiche

Table of Contents

Tommaso Agasisti, Patrizia Falzetti **pag. 2**
Socioeconomic sorting and test scores:an empirical analysis in the Italian junior secondary schools

Dario Albarello, Vera D'Amico **pag. 9**
Empirical testing of probabilistic seismic hazardmodels

Federico Andreis, Pier Alda Ferrari **pag. 15**
A proposal for the multidimensional extension ofCUB models

Morten Arendt Rasmussen, Evrim Acar **pag. 19**
Data fusion in the framework of coupled matrix tensor factorization withcommon, partially common and unique factors

Luigi Augugliaro, Angelo M. Mineo **pag. 20**
Estimation of Sparse Generalized LinearModels: the dglars package

Antonio Balzanella, Lidia Rivoli, Elvira Romano **pag. 24**
A comparison between two tools for data streamsummarization

Lucio Barabesi, Giancarlo Diana, Pier Francesco Perri **pag. 28**
Gini Index Estimation in Randomized ResponseSurveys

Francesco Bartolucci, Federico Belotti, Franco Peracchi **pag. 32**
A test for time-invariant individual effects ingeneralized linear models for panel data

Erich Battistin, Carlos Lamarche, Enrico Rettore **pag. 36**
Identification of the distribution of the causal effect of an intervention using a generalised factor model

Matilde Bini, Lucio Masserini **pag. 37**
Internal effectiveness of educational offer andstudents' satisfaction: a SEM approach

Matilde Bini, Leopoldo Nascia, Alessandro Zeli **pag. 41**
Groups heterogeneity and sectorsconcentration: a structural equation modelingfor micro level analysis of firms

Giuseppe Boari, Marta Nai Ruscone **pag. 45**
Use of Relevant Principal Components to Definea Simplified Multivarate Test Procedure ofOptimal Clutering

Giuseppe Boari, Gabriele Cantaluppi, Angelo Zanella **pag. 49**
Some Distance Proposals for Cluster Analysis inPresence of Ordinal Variables

Laura Bocci, Donatella Vicari **pag. 53**
A general model for INDCLUS with externalinformation

Paola Bongini, Paolo Trivellato, Mariangela Zenga **pag. 57**
The financial literacy and the undergraduates

Riccardo Bramante, Marta Nai Ruscone, Pasquale Spani **pag. 61**
Credit risk measurement and ethical issue: someevidences from the italian banks

Pierpaolo Brutti, Lucio Ceccarelli, Fulvio De Santis, Stefania Gubbiotti **pag. 65**
On the Stylometric Authorship of Ovid's DoubleHeroides: An Ensemble Clustering Approach

Silvia Caligaris, Fulvia Mecatti and Patrizia Farina **pag. 69**
Causal Inference in Gender Discrimination inChina: Nutrition, Health, Care

Giorgio Calzolari, Antonino Di Pino **pag. 73**
Self-Selection and Direct Estimation of Across-Regime Correlation Parameter

Maria Gabriella Campolo, Antonino Di Pino, Ester Lucia Rizzi **pag. 77**
Modern Vs. Traditional: A cluster-basedspecification of gender and familistic attitudesand their influence on the division of labour of Italian couples

Gabriele Cantaluppi, Marco Passarotti **pag. 81**
Clustering the Four Gospels in the Greek,Latin, Gothic and Old Church SlavonicTranslations

Carmela Cappelli, Francesca Di Iorio **pag. 85**
Regression Trees for change point analysis:methods, applications and recent developments

Roberto Casarin and Marco Tronzano and Domenico Sartore **pag. 89**
Bayesian Stochastic Correlation Models

Rosalia Castellano, Gennaro Punzo, Antonella Rocca **pag. 93**
Evaluating the selection effect in labour marketswith a low female participation

Paola Cerchiello, Paolo Giudici **pag. 97**
A statistical based H index for the evaluation ofe-markets

Annalisa Cerquetti **pag. 101**
Bayesian nonparametric estimation of globaldisclosure risk

Enrico Ciavolino, Roberto Savona **pag. 105**
The Forecasting side of Sovereign Risk: aGeneralized Cross Entropy Approach

Nicoletta Cibella, Tiziana Tuoto, Luca Valentino **pag. 109**
What data tell you that models can't say

Roberto Colombi, Sabrina Giordano **pag. 114**
Multiple Hidden Markov Models for CategoricalTime Series

Pier Luigi Conti, Daniela Marella **pag. 118**
Asymptotics in survey sampling for high entropysampling designs

Claudio Conversano, Massimo Cannas, Francessco Mola **pag. 122**
On the Use of Recursive Partitioning in CasualInference: A Proposal

Franca Crippa, Marcella Mazzoleni, Mariangela Zenga **pag. 128**
Keeping the pace with higher education. A fuzzystates gender study

F. Cugnata, C. Guglielmetti and S. Salini **pag. 133**
CUB model to validate FACIT TS-PSmeasurement instrument

Rosario D'Agata, Venera Tomaselli **pag. 137**
Multilevel Approach in Meta-Analysis of Pre-Election Poll Accuracy

Alfonso Iodice D'Enza and Angelos Markos **pag. 141**
Low-dimensional tracking of associationstructures in categorical data

Giulio D'Epifani **pag. 145**
Self-censored Categorical ResponsesA device for recovering latent behaviors

Pierpaolo D'Urso, Marta Disegna, Riccardo Massari **pag. 150**
Tourism Market Segmentation with ImpreciseInformation

Utkarsh J. Dang, Salvatore Ingrassia, Paul D. McNicholas and Ryan Browne **pag. 154**
Cluster-weighted models for multivariateresponse and extensions

Cristina Davino, Domenico Vistocco **pag. 158**
Unsupervised Classification through QuantileRegression

F. Marta L. Di Lascio, Simone Giannerini **pag. 162**
A copula-based approach to discoverinter-cluster dependence relationships

Josè G. Dias, Sofia B. Ramos **pag. 166**
Hierarchical market structure of Euro arearegime dynamics

Drago Carlo, Balzanella Antonio **pag. 170**
Consensus Community Detection: a NonmetricMDS Approach

Fabrizio Durante, Roberta Pappad`a and Nicola Torelli **pag. 174**
Clustering financial time series by measures oftail dependence

Marco Enea, Antonella Plaia **pag. 178**
Influence diagnostics for generalized linearmixed models: a gradient-like statistic

Enrico Fabrizi, Maria R. Ferrante, Carlo Trivisano **pag. 182**
Joint estimation of poverty and inequalityparameters in small areas

Giorgio Fagiolo, Andrea Roventini **pag. 187**
Macroeconomic Policy in DSGE andAgent-Based Models

Salvatore Fasola, Mariangela Sciandra **pag. 191**
New Flexible Probability Distributions forRanking Data

Maria Brigida Ferraro, Paolo Giordani **pag. 195**
A new fuzzy clustering algorithm with entropyregularization

Camilla Ferretti, Piero Ganugi, Renato Pieri **pag. 199**
Mobility measures for the dairy farms inLombardy

Silvia Figini, Marika Vezzoli **pag. 203**
Model averaging and ensemble methods for riskcorporate estimation

Luis Angel García-Escudero, Alfonso Gordaliza, Carlos Matrán, Agustín Mayo-Iscar **pag. 207**
New proposals for clustering based on trimmingand restrictions

Andreas Geyer-Schulz, Fabian Ball **pag. 211**
Formal Diagnostics for Graph Clustering: TheRole of Graph Automorphisms

Massimiliano Giacalone, Angela Alibrandi **pag. 215**
An overview on multiple regression models basedon permutation tests

Francesca Giambona, Mariano Porcu **pag. 219**
The determinants of Italian students' readingscores: a Quantile Regression analysis

Paolo Giordani, Henk A.L. Kiers, Maria Antonietta Del Ferraro **pag. 223**
The R Package ThreeWay

Giuseppe Giordano, Ilaria Primerano **pag. 227**
Co-occurence Network from SemanticDifferential Data

Paolo Giudici **pag. 231**
Financial risk data analysis

Silvia Golia, Anna Simonetto **pag. 237**
A Comparison between SEM and Rasch model:the polytomous case

Anna Gottard **pag. 241**
Some considerations on VCUB models

Francesca Greselin, Salvatore Ingrassia **pag. 245**
Data driven EM constraints for mixtures offactor analyzers

Leonardo Grilli, Carla Rampichini, Roberta Varriale **pag. 249**
Predicting students' academic performance: achallenging issue in statistical modelling

Luigi Grossi, Fany Nan **pag. 255**
Robust estimation of regime switching models

Kristian Hovde Liland **pag. 259**
Variable selection in sequential multi-block analysis

Maria Iannario **pag. 260**
Robustness issues for a class of models forordinal data

Maria Iannario, Domenico Piccolo **pag. 264**
A class of ordinal data models in R

Salvatore Ingrassia, Antonio Punzo **pag. 268**
Parsimony in Mixtures with Random Covariates

Hiroshi Inoue **pag. 272**
International Relations Based on the VotingBehavior in General Assembly

Carmela Iorio, Massimo Aria, Antonio D'Ambrosio **pag. 276**
Visual model representation and selection forclassification and regression trees

Monia Lupparelli, Luca La Rocca, Alberto Roverato **pag. 284**
Log-Mean Linear Parameterizations for SmoothIndependence Models

Marica Manisera, Paola Zuccolotto **pag. 288**
Nonlinear CUB models

Marica Manisera, Marika Vezzoli **pag. 292**
Finding number of groups using a penalizedinternal cluster quality index

Daniela Marella, Paola Vicard **pag. 296**
Object-Oriented Bayesian Network to deal withmeasurement error in household surveys

Angelos Markos, Alfonso Iodice D'Enza, Michel Van de Velden **pag. 300**
Beyond tandem analysis: joint dimensionreduction and clustering in R

F. Martella and M. Alfò **pag. 304**
A biclustering approach for discrete outcomes

Mariagiulia Matteucci, Stefania Mignani, Roberto Ricci **pag. 309**
A Multidimensional IRT approach to analyzelearning achievement of Italian students

Sabina Mazza **pag. 314**
Extending the Forward Search to theCombination of Multiple Classifiers: A Proposal

Fulvia Mecatti, M. Giovanna Ranalli **pag. 318**
Plug-in Bootstrap for Sample Survey Data

Alessandra Menafoglio, Matilde Dalla Rosa and Piercesare Secchi **pag. 322**
A BLU Predictor for Spatially DependentFunctional Data of a Hilbert Space

Maria Adele Milioli, Lara Berzieri, Sergio Zani **pag. 326**
Comparing fuzzy and multidimensional methodsto evaluate well-being at regional level

Michelangelo Misuraca, Maria Spano **pag. 331**
Comparing text clustering algorithms from amultivariate perspective

Cristina Mollica, Luca Tardella **pag. 335**
Mixture models for ranked data classification

Isabella Morlini, Stefano Orlandini **pag. 339**
Cluster analysis of three-way atmospheric data

Roberto Nardecchia, Roberto Sanzo, Margherita Velucchi, Alessandro Zeli **pag. 345**
Productivity transition probabilities: A microlevel data analysis for Italian manufacturingsectors (1998-2007)

Andrea Neri, Giuseppe Ilardi **pag. 349**
Interviewers, co-operation and data accuracy: isthere a link?

Akinori Okada, Satoru Yokoyama **pag. 353**
Nonhierarchical Asymmetric Cluster Analysis

Marco Perone Pacifico **pag. 357**
SuRF: Subspace Ridge Finder

Andrea Pagano, Francesca Torti, Jessica Cariboni, Domenico Perrotta **pag. 361**
Robust clustering of EU banking data

Giuseppe Pandolfo, Giovanni C. Porzio **pag. 365**
On depth functions for directional data

Andrea Pastore, Stefano F. Tonellato **pag. 369**
A generalised Silhouette-width measure

Fulvia Pennoni, Giorgio Vittadini **pag. 373**
Hospital efficiency under two competing paneldata models

Alessia Pini, Simone Vantini **pag. 377**
The Interval-Wise Control of the Family-WiseError Rate for Testing Functional Data

Mariano Porcu, Isabella Sulis **pag. 381**
Detecting differences between primary schools inmathematics and reading achievement by usingschools added-value measures of performance

Antonio Punzo, Paul D. McNicholas, Katherine Morris, Ryan P. Browne **pag. 387**
Outlier Detection via Contaminated MixtureDistributions

Emanuela Raffinetti, Pier Alda Ferrari **pag. 392**
New perspectives for the RDI index in socialresearch fields

Monia Ranalli, Roberto Rocci **pag. 396**
Mixture models for ordinal data: a pairwiselikelihood approach

Marco Riani, Andrea Cerioli, Gianluca Morelli **pag. 400**
Issues in robust clustering

Stèphane Robin **pag. 404**
Deciphering and modeling heterogeneity ininteraction networks

Rosaria Romano, Francesco Palumbo **pag. 409**
Partial Possibilistic Regression Path Modeling

Renata Rotondi **pag. 413**
Classsification of composite seismogenic sourcesthrough probabilitic score indices

Gabriella Schoier **pag. 417**
On Wild Bootstrap and M Unit Root Test

Luca Scrucca **pag. 421**
On the implementation of a parallel algorithmfor variable selection in model-based clustering

Paolo Sestito **pag. 425**
The Role of Learning Measurement in theGovernance of an Education System: anOverview of the Issues

John Shawe-Taylor, Blaz Zlicar **pag. 430**
Novelty Detection with Support Vector Machines

Nadia Solaro **pag. 431**
Multidimensional scaling with incompletedistance matrices: an insight into the problem

Luigi Spezia, Cecilia Pinto **pag. 435**
Markov switching models for high-frequencytime series: flapper skate's depth profile as a casestudy

Ralf Stecking, Klaus B. Schebesch **pag. 439**
Data Privacy in Credit Scoring: Evaluating SVMApproaches Based on Microaggregated Data

Isabella Sulis, Francesca Giambona, Nicola Tedesco **pag. 443**
Analyzing university students' careers usingMulti-State Models

Luca Tardella, Danilo Alunni Fegatelli **pag. 447**
BBRecap for Bayesian BehaviouralCapture-Recapture Modeling

Cristina Tortora, Paul D. McNicholas, Ryan P. Browne **pag. 451**
Mixtures of generalized hyperbolic factoranalyzers

Giovanni Trovato **pag. 455**
Testing for endogeneity and countryheterogeneity

Joaquin Vanschoren and Mikio L. Braun, Cheng Soon Ong **pag. 461**
Open science in machine learning

Valerio Veglio **pag. 465**
Logistic Regression and Decision Tree:Performance Comparisons in EstimatingCustomers' Risk of Churn

Maurizio Vichi **pag. 469**
Robust Two-mode clustering

Vincenzina Vitale **pag. 470**
Hierarchical Graphical Models and ItemResponse Theory

Sara Viviani **pag. 474**
Extending the JM libraRy

Adalbert F.X. Wilhelm **pag. 478**
Visualisations of Classification Tree Models: AnEvaluative Comparison

On the Stylometric Authorship of Ovid's Double Heroides: An Ensemble Clustering Approach

Pierpaolo Brutti, Lucio Ceccarelli, Fulvio De Santis, Stefania Gubbiotti

Abstract *Double Heroides* are six elegies traditionally attributed to Ovid, whose authenticity have been repeatedly questioned. As a contribution to establish the period of composition of these elegies, this article proposes a statistical analysis based on consensus clustering of mixed data composed of standard (i.e. scalar) and compositional variables derived by an extensive metrical study of the Ovid's poetic production.

Key words: Consensus clustering, compositional data, Latin metric.

1 Introduction: Ovid's double letters

A relevant open problem in Latin literature is the authorship and the date of Ovids *Heroides XVI-XXI* (*double Heroides*). The *Heroides* are a collection of 21 elegies written in *elegiac couplets*. The authenticity of the collection has been questioned since Lackmann [1] who, on the basis of certain metrical anomalies with respect to genuine Ovidian poems as well as of non metrical considerations, claimed that some of the *Heroides* (including the double Heroides) where not composed by Ovid. In this article we propose a statistical approach to provide further elements of discussion on the dating problem. In particular our attempt is to find statistical support, based on metrical features, to the hypotheses that *double Heroides* belong to one of the three main phases of Ovid's poetic production: the early Roman period, the mature Roman period and the exile period. The stylometric methodology we adopt is based on consensus clustering techniques for mixed data composed of standard (i.e. scalar) and compositional variables.

Lucio Ceccarelli
Università degli Studi dell'Aquila e-mail: lucio.ceccarelli@univaq.it

Pierpaolo Brutti, Fulvio De Santis, Stefania Gubbiotti
Sapienza Università di Roma, e-mail: stefania.gubbiotti@uniroma1.it

2 Basics of metric: the elegiac couplet

Meter is the basic rhythmic structure of a verse. The study of meters and forms of versification is known as metric. We here focus on the so called *elegiac couplet* or *distich*, a poetic form initially introduced in Greek lyric, later on adopted by Roman poets, and in particular by Ovid. To illustrate the "anatomy" of the elegiac couplet, let us consider one of the 406 distichs from Ovid's *Fasti VI*: The main steps in the

Tempora labuntur, tacitisque senescimus annis;
et fugiunt freno non remorante dies.

metrical analysis are summarized as follows (see Figure 2).

- Each syllable of the words of a verse is categorized as *long* (−) or *short* (∪) according to its *weight*. (e.g. in Tempora, Tem- is long, -po- and -ra are short).
- Specific sequences of syllables define a *foot* (delimited by $|\ldots|$), for instance

Dactyl (D)	formed by 3 syllables	− ∪ ∪
Spondee (S)	formed by 2 syllables	− −
Trochee (T)	formed by 2 syllables	− ∪

- Each verse is formed by a specific number of *feet*. The first verse of the elegiac distich is called *hexameter*; the second one is called *pentameter*.
- The metrical pattern of feet in a verse is then summarized by a sequence of letters. For instance, in the couplet above the scheme is DSDDDS · DD−DD−

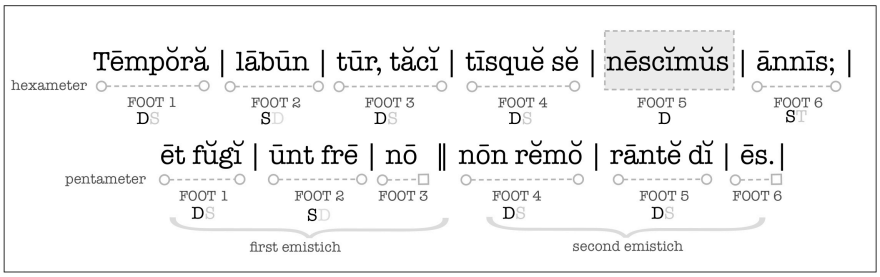

Fig. 1 "Anatomy" of the elegiac couplet; (−) and (∪) denote long and short syllables respectively.

In general, in the hexameter each of the first four feet can be alternatively a *dactyl* or a *spondee*. The fifth foot is almost always a *dactyl*. The sixth foot is either a *spondee* or a *trochee*. Conversely, the pentamer is made up of two equal parts containing two dactyls followed by a long syllable (*hemistich*). Spondees replace dactyls in the first half, but never in the second. The choice of a particular scheme for the verses, together with many other metrical features that will be mentioned in the following section, yield a great variability in the realization of the elegiac distich, that strongly characterizes the style of each single author. For instance, if we consider the first four feet of the hexameter only, there are $2^4 = 16$ possible alternative choices of dactyls and spondees. In summary, the metrical technique is

a very personal ability of the poet and it reflects not only his skills, but also his sensitivity. In this sense, the stylometric analysis can be helpful in the attempt of attributing a poem to a specific author.

3 Stylometric features

The goal of a metric study is to identify characterizing stylistic features of a poet with respect to the metric language of the tradition, as well as to detect deviations of metric features of some parts of his own poetic production with respect to his entire work. This kind of study requires a translation of metrical phenomena into quantitative data, whose relevance can be pointed out uniquely by appropriate statistical analysis of poetic *corpora* (see [2]). All the poetic production in elegiac couplets attributed to Ovid has been examined from a metrical point of view in [3]. In this metrical study many quantitative informations has been extracted and collected on each section of every poem, such as the total frequency of the dactyls and the distribution of dactyls both in the first four feet of the hexameter and in the first half of the pentameter, the presence of short syllables in the couplet, the occurrence of some particular forms (e.g. *synalepha*, *clausula*), and so on. Consequently, our dataset consists of 15 distinct metrical features measured on each of the 27 poems considered as statistical units. Most of these 15 variables are actually frequency distributions summarizing the occurrence of a specific metrical phenomenon over a whole poem, and therefore they will be treated as *compositional data* (see [1] and [8]). For example, in Figure 3 are represented three variables observed on two of

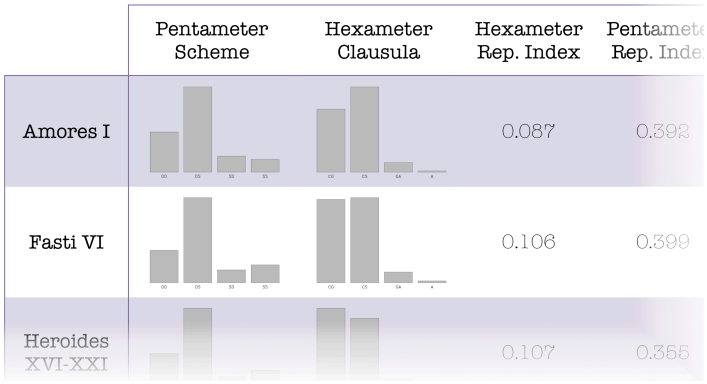

Fig. 2 A snippet of the stylometric dataset.

the poems: the first and the second one consist of distributions associated to the realizations of the pentameter scheme, and to the clausula in the exameter respectively, whereas the third scalar one is a particular stylometric quantity usually called *replication index*.

4 Methods and results

Consensus clustering, emerged as an important elaboration of the classical clustering problem. These methods are commonly used to establish consensus among multiple clustering algorithms, or multiple realizations of the same clustering algorithm on a single dataset, or even to integrate multi–source data (for a survey see [7]). In this work we adapt to the peculiarities of the data at hand, two recently proposed tecniques, namely the *Multiple Dataset Integration* method developed in [4] and the Bayesian consensus clustering based on finite Dirichlet mixture models described in [6]. Both approaches use a statistical framework to cluster each data source separately while simultaneously modeling dependence between the clusterings in order to borrow strength across data sources. Aggregation mechanisms of this type make the resulting overall clustering more robust and stable of other unsupervised classification solutions, while still allowing to better interpret the contribution of each data–source to the final partition.

In summary, the consensus clustering shows a quite remarkable distinction between three groups, characterized by different metric profiles. It can be noticed that in each group one of the three periods prevails. Finally, the *double Heroides* turn out to be compatible with the (stylo)metric features of the latest period of Ovid's poetry.

References

1. Aitchison J.: The Statistical Analysis of Compositional Data. The Blackburn Press, (2003)
2. Ceccarelli L.: Contributi sulla storia dell'esametro latino. Herder, Roma (2008)
3. Ceccarelli L.: L'evoluzione del distico elegiaco tra Catullo e Ovidio. In: Cristofoli R., Santini C., Santucci F. Properzio tra tradizione e innovazione, Atti del convegno internazionale (Assisi-Spello, 21-23 maggio 2010), Assisi 2012, pp. 47–97.
4. Kirk P., Griffin J.E., Savage R.S., Ghahramani, Z., Wild, D.L.: Bayesian correlated clustering to integrate multiple datasets. Bioinformatics, **28**(24), 3290-3297, (2012)
5. Lachmann K.: De Ovidi epistulis. Progr. Univers. Berolinensis, (1848)
6. Lock E.F., Dunson D.B.: Bayesian consensus clustering. arXiv:1302.7280, (2013)
7. Nguyen, N., Caruana, R. Consensus clusterings. In: Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), October 28-31, 2007, Omaha, Nebraska, USA, pp. 607-612. IEEE Computer Society, 2007
8. van den Boogaart K.G., Tolosana-Delgado R.: Analyzing Compositional Data with R. Springer, (2013)