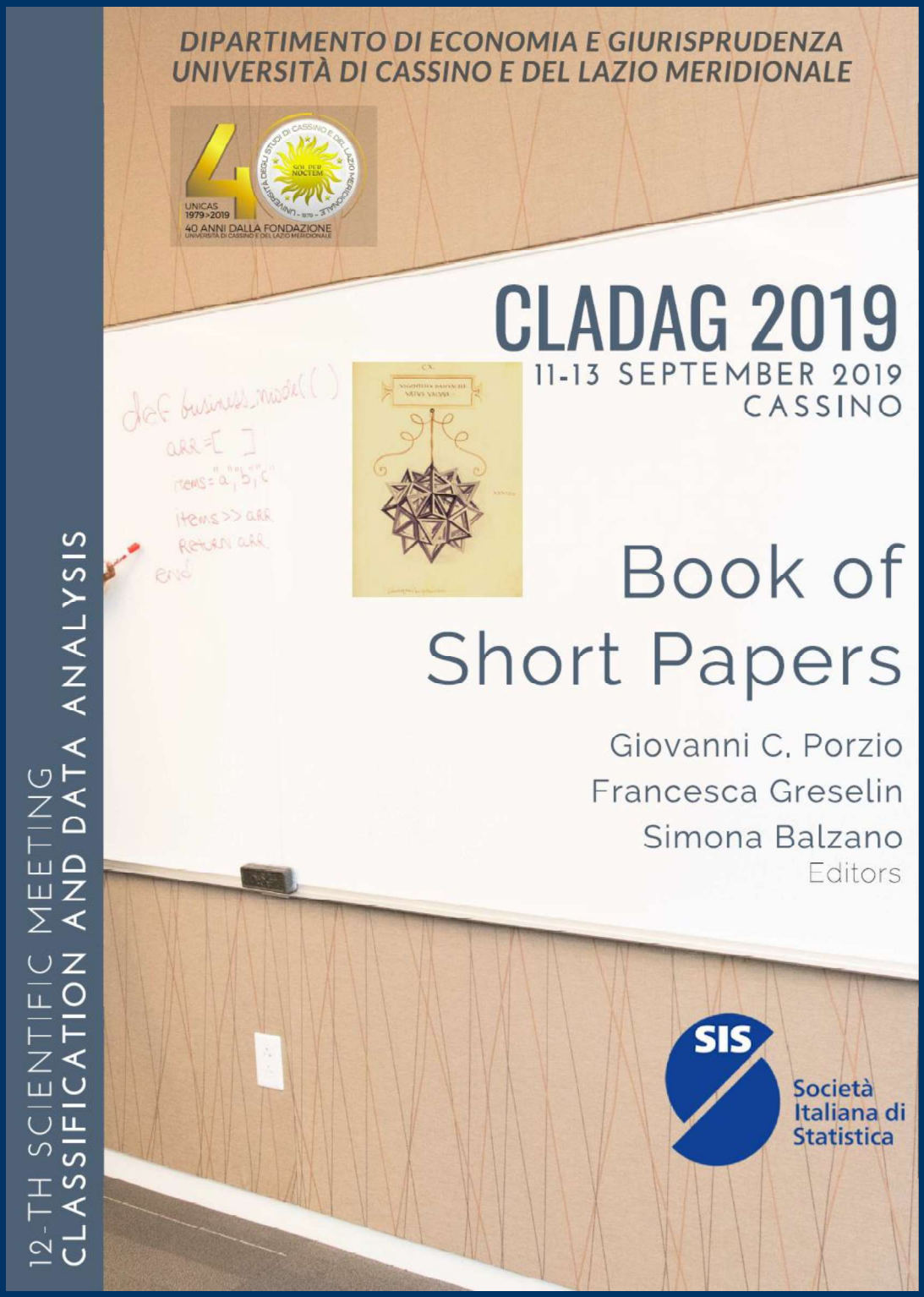
UNIVERSITÀ degli STUDI di CASSINO e del LAZIO MERIDIONALE

COLLANA SCIENTIFICA



EDIZIONI UNIVERSITA' DI CASSINO

2019

Università di Cassino e del Lazio Meridionale
Centro Editoriale di Ateneo
Palazzo degli Studi Località Folcara, Cassino (FR), Italia

ISBN 978-88-8317-108-6

CLADAG 2019

Book of Short Papers

Giovanni C. Porzio

Francesca Greselin

Simona Balzano

Editors

2019

Contents

Keynotes lectures

Unifying data units and models in (co-)clustering 3
Christophe Biernacki

Statistics with a human face 4
Adrian Bowman

Bayesian model-based clustering with flexible and sparse priors 5
Bettina Grün

Grinding massive information into feasible statistics: current challenges and opportunities for data scientists 6
Francesco Mola

Statistical challenges in the analysis of complex responses in biomedicine 7
Sylvia Richardson

Invited and contributed sessions

Model-based clustering of time series data: a flexible approach using nonparametric state-switching quantile regression models 8
Timo Adam, Roland Langrock, Thomas Kneib

Some issues in generalized linear modeling 12
Alan Agresti

Assessing social interest in burnout using functional data analysis through google trends 16
Ana M. Aguilera, Francesca Fortuna, Manuel Escabias

Measuring equitable and sustainable well-being in Italian regions: a non-aggregative approach 20
Leonardo Salvatore Alaimo, Filomena Maggino

Bootstrap inference for missing data reconstruction 22
Giuseppina Albano, Michele La Rocca, Maria Lucia Parrella, Cira Perna

Archetypal contour shapes 26
Aleix Alcacer, Irene Epifanio, M. Victoria Ibáñez, Amelia Simó

Random projections of variables and units 30
Laura Anderlucci, Roberta Falcone, Angela Montanari

Sparse linear regression via random projections ensembles 34
Laura Anderlucci, Matteo Farnè, Giuliano Galimberti, Angela Montanari

High-dimensional model-based clustering via random projections 38
Laura Anderlucci, Francesca Fortunato, Angela Montanari

Multivariate outlier detection in high reliability standards fields using ICS 42
Aurore Archimbaud, Klaus Nordhausen, Anne Ruiz-Gazen

Evaluating the school effect: adjusting for pre-test or using gain scores? 45
Bruno Arpino, Silvia Bacci, Leonardo Grilli, Raffaele Guetto, Carla Rampichini

ACE, AVAS and robust data transformations 49
Anthony Atkinson

Mixtures of multivariate leptokurtic Normal distributions 53
Luca Bagnato, Antonio Punzo, Maria Grazia Zoia

Detecting and interpreting the consensus ranking based on the weighted Kemeny distance 57
Alessio Baldassarre, Claudio Conversano, Antonio D'Ambrosio

Predictive principal components analysis 61
Simona Balzano, Maja Bozic, Laura Marcis, Renato Salvatore

Flexible model-based trees for count data 63
Federico Banchelli

Euclidean distance as a measure of conformity to Benford's law in digital analysis for fraud detection 67
Mateusz Baryła, Józef Pociecha

The evolution of the purchase behavior of sparkling wines in the Italian market 71
Francesca Bassi, Fulvia Pennoni, Luca Rossetto

Modern likelihood-frequentist inference at work 75
Ruggero Bellio, Donald A. Pierce

Ontology-based classification of multilingual corpuses of documents 79
Sergey Belov, Salvatore Ingrassia, Zoran Kalinić, Paweł Lula

Modeling heterogeneity in clustered data using recursive partitioning 83
Moritz Berger, Gerhard Tutz

Mixtures of experts with flexible concomitant covariate effects: a bayesian solution 87
Marco Berrettini, Giuliano Galimberti, Thomas Brendan Murphy, Saverio Ranciati

Sampling properties of an ordinal measure of interrater absolute agreement 91
Giuseppe Bove, Pier Luigi Conti, Daniela Marella

Tensor analysis can give better insight 95
Rasmus Bro

A boxplot for spherical data 97
Davide Buttarazzi, Giuseppe Pandolfo, Giovanni C. Porzio, Christophe Ley

Machine learning models for forecasting stock trends 99
Giacomo Camba, Claudio Conversano

Tree modeling ordinal responses: CUBREMOT and its applications 103
Carmela Cappelli, Rosaria Simone, Francesca Di Iorio

Supervised learning in presence of outliers, label noise and unobserved classes 104
Andrea Cappozzo, Francesca Greselin, Thomas Brendan Murphy

Asymptotics for bandwidth selection in nonparametric clustering 108
Alessandro Casa, José E. Chacón, Giovanna Menardi

Foreign immigration and pull factors in Italy: a spatial approach 112
Oliviero Casacchia, Luisa Natale, Francesco Giovanni Truglia

Dimensionality reduction via hierarchical factorial structure 116
Carlo Cavicchia, Maurizio Vichi, Giorgia Zaccaria

Likelihood-type methods for comparing clustering solutions 120
Luca Coraggio, Pietro Coretto

Labour market analysis through transformations and robust multilevel models 124
Aldo Corbellini, Marco Magnani, Gianluca Morelli

Modelling consumers' qualitative perceptions of inflation 128
Marcella Corduas, Rosaria Simone, Domenico Piccolo

Noise resistant clustering of high-dimensional gene expression data 132
Pietro Coretto, Angela Serra, Roberto Tagliaferri

Classify X-ray images using convolutional neural networks 136
Federica Crobu, Agostino Di Ciaccio

A compositional analysis approach assessing the spatial distribution of trees in Guadalajara, Mexico 140
Marco Antonio Cruz, Maribel Ortego, Elisabet Roca

Joining factorial methods and blockmodeling for the analysis of affiliation networks 142
Daniela D'Ambrosio, Marco Serino, Giancarlo Ragozini

A latent space model for clustering in multiplex data 146
Silvia D'Angelo, Michael Fop

Post processing of two dimensional road profiles: variogram scheme application and sectioning procedure 150
Mauro D'Apuzzo, Rose-Line Spacagna, Azzurra Evangelisti, Daniela Santilli, Vittorio Nicolosi

A new approach to preference mapping through quantile regression 154
Cristina Davino, Tormod Naes, Rosaria Romano, Domenico Vistocco

On the robustness of the cosine distribution depth classifier 158
Houyem Demni, Amor Messaoud, Giovanni C. Porzio

Network effect on individual scientific performance: a longitudinal study on an Italian scientific community 162
Domenico De Stefano, Giuseppe Giordano, Susanna Zaccarin

Penalized vs constrained maximum likelihood approaches for clusterwise linear regression modelling 166
Roberto Di Mari, Stefano Antonio Gattone, Roberto Rocci

Local fitting of angular variables observed with error 170
Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor

Quantile composite-based path modeling to estimate the conditional quantiles of health indicators 174
Pasquale Dolce, Cristina Davino, Stefania Taralli, Domenico Vistocco

AUC-based gradient boosting for imbalanced classification 178
Martina Dossi, Giovanna Menardi

How to measure material deprivation? A latent Markov model based approach 182
Francesco Dotto

Decomposition of the interval based composite indicators by means of biclustering 186
Carlo Drago

Consensus clustering via pivotal methods 190
Leonardo Egidi, Roberta Pappadà, Francesco Pauli, Nicola Torelli

Robust model-based clustering with mild and gross outliers 194
Alessio Farcomeni, Antonio Punzo

Gaussian processes for curve prediction and classification 198
Sara Fontanella, Lara Fontanella, Rosalba Ignaccolo, Luigi Ippoliti, Pasquale Valentini

A new proposal for building immigrant integration composite indicator 199
Venera Tomaselli, Mario Fordellone, Maurizio Vichi

Biodiversity spatial clustering 203
Francesca Fortuna, Fabrizio Maturo, Tonio Di Battista

Skewed distributions or transformations? Incorporating skewness in a cluster analysis 207
Michael Gallaugher, Paul McNicholas, Volodymyr Melnykov, Xuwen Zhu

Robust parsimonious clustering models 208
Luis Angel Garcia-Escudero, Agustin Mayo-Iscar, Marco Riani

Projection-based uniformity tests for directional data 212
Eduardo García-Portugués, Paula Navarro-Esteban, Juan Antonio Cuesta-Albertos

Graph-based clustering of visitors' trajectories at exhibitions 214
Martina Gentilin, Pietro Lovato, Gloria Menegaz, Marco Cristani, Marco Minozzo

Symmetry in graph clustering 218
Andreas Geyer-Schulz, Fabian Ball

Bayesian networks for the analysis of entrepreneurial microcredit: evidence from Italy 222
Lorenzo Giammei, Paola Vicard

The PARAFAC model in the maximum likelihood approach 226
Paolo Giordani, Roberto Rocci, Giuseppe Bove

Structure discovering in nonparametric regression by the GRID procedure 230
Francesco Giordano, Soumendra Nath Lahiri, Maria Lucia Parrella

A microblog auxiliary part-of-speech tagger based on bayesian networks 234
Silvia Golia, Paola Zola

Recent advances in model-based clustering of high dimensional data 238
Isobel Claire Gormley

Tree embedded linear mixed models 239
Anna Gottard, Leonardo Grilli, Carla Rampichini, Giulia Vannucci

Weighted likelihood estimation of mixtures 243
Luca Greco, Claudio Agostinelli

A canonical representation for multiblock methods 247
Mohamed Hanafi

An adequacy approach to estimating the number of clusters 251
Christian Hennig

Classification with weighted compositions 255
Karel Hron, Julie Rendlova, Peter Filzmoser

MacroPCA: an all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers 256
Mia Hubert, Peter J. Rousseeuw, Wannes Van den Bossche

Marginal effects for comparing groups in regression models for ordinal outcome when uncertainty is present 258
Maria Iannario, Claudia Tarantola

A multi-criteria approach in a financial portfolio selection framework 262
Carmela Iorio, Giuseppe Pandolfo, Roberta Siciliano

Clustering of trajectories using adaptive distances and warping 266
Antonio Irpino, Antonio Balzanella

Sampling and learning Mallows and generalized Mallows models under the Cayley distance: short paper 270
Ekhine Irurozki, Borja Calvo, Jose A. Lozano

The gender parity index for the academic students progress 274
Aglaia Kalamatianou, Adele H. Marshall, Mariangela Zenga

Some asymptotic properties of model selection criteria in the latent block model 278
Christine Keribin

Invariant concept classes for transcriptome classification 282
Hans Kestler, Robin Szekely, Attila Klimmek, Ludwig Lausser

Clustering of ties defined as symbolic data 283
Luka Kronegger

Application of data mining in the housing affordability analysis 284
Viera Labudová, Ľubica Sipková

Cylindrical hidden Markov fields 288
Francesco Lagona

Comparing tree kernels performances in argumentative evidence classification 292
Davide Liga

Recent advancement in neural network analysis of biomedical big data 296
Pietro Liò, Giovanna Maria Dimitri, Chiara Sopegno

Bias reduction for estimating functions and pseudolikelihoods 297
Nicola Lunardon

Large scale social and multilayer networks 301
Matteo Magnani

Uncertainty in statistical matching by BNs 305
Daniela Marella, Paola Vicard, Vincenzina Vitale

Evaluating the recruiters' gender bias in graduate competencies 309
Paolo Mariani, Andrea Marletta

Dynamic clustering of network data: a hybrid maximum likelihood approach 313
Maria Francesca Marino, Silvia Pandolfi

Stability of joint dimension reduction and clustering 317
Angelos Markos, Michel Van de Velden, Alfonso Iodice D'Enza

Hidden Markov models for clustering functional data 321
Andrea Martino, Giuseppina Guatteri, Anna Maria Paganoni

Composite likelihood inference for simultaneous clustering and dimensionality reduction of mixed-type longitudinal data 325
Antonello Maruotti, Monia Ranalli, Roberto Rocci

Bivariate semi-parametric mixed-effects models for classifying the effects of Italian classes on multiple student achievements 329
Chiara Masci, Francesca Ieva, Tommaso Agasisti, Anna Maria Paganoni

Multivariate change-point analysis for climate time series 333
Gianluca Mastrantonio, Giovanna Jona Lasinio, Alessio Pollice, Giulia Capotorti, Lorenzo Teodonio, Carlo Blasi

A dynamic stochastic block model for longitudinal networks 337
Catherine Matias, Tabea Rebafka, Fanny Villers

Unsupervised fuzzy classification for detecting similar functional objects 339
Fabrizio Maturo, Francesca Fortuna, Tonio Di Battista

Mixture modelling with skew-symmetric component distributions 343
Geoffrey McLachlan

New developments in applications of pairwise overlap 344
Volodymyr Melnykov, Yana Melnykov, Domenico Perrotta, Marco Riani, Francesca Torti, Yang Wang

Modelling unobserved heterogeneity of ranking data with the bayesian mixture of extended Plackett-Luce models 346
Cristina Mollica, Luca Tardella

Issues in nonlinear time series modeling of European import volumes 350
Gianluca Morelli, Francesca Torti

Gaussian parsimonious clustering models with covariates and a noise component 352
Keefe Murphy, Thomas Brendan Murphy

Illumination in depth analysis 353
Stanislav Nagy, Jiří Dvořák

Copula-based non-metric unfolding on augmented data matrix 357
Marta Nai Ruscone, Antonio D'Ambrosio

A statistical model for software releases complexity prediction 361
Marco Ortu, Giuseppe Destefanis, Roberto Tonelli

Comparison of serious diseases mortality in regions of V4 365
Viera Pacáková, Lucie Kopecká

Price and product design strategies for manufacturers of electric vehicle batteries: inferences from latent class analysis 369
Friederike Paetz

A Mahalanobis-like distance for cylindrical data 373
Lucio Palazzo, Giovanni C. Porzio, Giuseppe Pandolfo

Archetypes, prototypes and other types 377
Francesco Palumbo, Giancarlo Ragozini, Domenico Vistocco

Generalizing the skew-t model using copulas 381
Antonio Parisi, Brunero Liseo

Contamination and manipulation of trade data: the two faces of customs fraud 385
Domenico Perrotta, Andrea Cerasa, Lucio Barabesi, Mario Menegatti, Andrea Cerioli

Bayesian clustering using non-negative matrix factorization 389
Michael Porter, Ketong Wang

Exploring gender gap in international mobility flows through a network analysis approach 393
Ilaria Primerano, Marialuisa Restaino

Clustering two-mode binary network data with overlapping mixture model and covariates information 395
Saverio Ranciati, Veronica Vinciotti, Ernst C. Wit, Giuliano Galimberti

A stochastic blockmodel for network interaction lengths over continuous time 399
Riccardo Rastelli, Michael Fop

Computationally efficient inference for latent position network models 403
Riccardo Rastelli, Florian Maire, Nial Friel

Clustering of complex data stream based on barycentric coordinates 407
Parisa Rastin, Basarab Matei, Guénaël Cabanes

An INDSCAL based mixture model to cluster mixed-type of data 411
Roberto Rocci, Monia Ranalli

Topological stochastic neighbor embedding 415
Nicoleta Rogovschi, Nistor Grozavu, Basarab Matei, Younès Bennani, Seiichi Ozawa

Functional data analysis for spatial aggregated point patterns in seismic science 419
Elvira Romano, Jonatan González Monsalve, Francisco Javier Rodríguez Cortés, Jorge Mateu

ROC curves with binary multivariate data 420
Lidia Sacchetto, Mauro Gasparini

Silhouette-based method for portfolio selection 424
Marco Scaglione, Carmela Iorio, Antonio D'Ambrosio

Item weighted Kemeny distance for preference data 428
Mariangela Sciandra, Simona Buscemi, Antonella Plaia

A fast and efficient modal EM algorithm for Gaussian mixtures 432
Luca Scrucca

Probabilistic archetypal analysis 436
Sohan Seth

Multilinear tests of association between networks 438
Daniel K. Sewell

Use of multi-state models to maximise information in pressure ulcer prevention trials 442
Linda Sharples, Isabelle Smith, Jane Nixon

Partial least squares for compositional canonical correlation 445
Violetta Simonacci Massimo Guarino, Michele Gallo

Dynamic modelling of price expectations 449
Rosaria Simone, Domenico Piccolo, Marcella Corduas

Towards axioms for hierarchical clustering of measures 453
Philipp Thomann, Ingo Steinwart, Nico Schmid

Influence of outliers on cluster correspondence analysis 454
Michel Van de Velden, Alfonso Iodice D'Enza, Lisa Schut

Earthquake clustering and centrality measures 458
Elisa Varini, Antonella Peresan, Jiancang Zhuang

Co-clustering high dimensional temporal sequences summarized by histograms 462
Rosanna Verde, Antonio Irpino, Antonio Balzanella

Statistical analysis of item pre-knowledge in educational tests: latent variable modelling and optimal statistical decision 466
Chen Yunxiao, Lu Yan, Irini Moustaki

Evaluation of the web usability of the University of Cagliari portal: an eye tracking study 468
Gianpaolo Zammarchi, Francesco Mola

Application of survival analysis to critical illness insurance data 472
David Zapletal, Lucie Kopecka

A NEW PROPOSAL FOR BUILDING IMMIGRANT INTEGRATION COMPOSITE INDICATOR

Venera Tomaselli[1], Mario Fordellone[2] and Maurizio Vichi[2]

[1] Department of Political and Social Sciences, University of Catania,
(e-mail: venera.tomaselli@unict.it)

[2] Department of Statistical Sciences, Sapienza University of Rome,
(e-mail: mario.fordellone@uniroma1.it, maurizio.vichi@uniroma1.it)

ABSTRACT: Integration consists in a multidimensional process, which can take place in different ways and in different times in relation to each single economic, social, cultural, and political dimension. In this paper, we aim at providing a methodological proposal based on PLS-SEM to build a composite immigrant integration indicator.

KEYWORDS: partial least squares, immigrant integration, composite indicator, structural equation modelling.

1 Measuring immigrant integration

Integration consists in a multidimensional process, which can take place in different ways and in different times in relation to each single economic, social, cultural, and political dimension. It aims at pursuing mutual respect of ethno-cultural differences and peaceful coexistence among populations within a historical and social reality. Its goal cannot be reached once for all but must be continuously pursued distinguishing different integration processes at economic, cultural, social, and political level. A high economic integration level may be quickly achieved, indeed, along with scarce or no social or political integration. Each single dimension, diachronically positioned over time, generates different integration levels. Hence, examining each single dimension is important as well as building composite indexes simultaneously comprehensive of all dimensions in order to obtain a full description of a complex phenomenon and to convey a suitable set of information.

According to the literature (Entzinger, 2000), the concept of integration can be broken down into different dimensions. Firstly, the socio-economic dimension refers to housing conditions, work conditions and income. Including mostly the theme of citizenship, also the legal-political dimension takes into account two sub-dimensions. The other sub-dimension concerns the rights of political participation - from the freedom of association to the voting right - which in some countries can be used at local government elections even without having achieved the citizenship status of the host country. Finally, the cultural and social dimension considers

several elements, among which knowledge of the Italian language, free times activities and access to information.

In this paper, we aim at providing a methodological proposal to build a composite immigrant integration indicator, able to measure the different aspects related to integration, such as employment, education, social inclusion, and active citizenship. With this in mind, we analyse the data collected from European Social Survey (ESS), Round 8, on immigration by the Partial Least Squares Path Modelling (PLSPM) approach (Tenenhaus et al., 2005). The PLSPM models are Structural Equation Modelling suitable to estimate interaction and main effects among multiple sets of latent variables. In the present study we use a simultaneous non-hierarchical clustering and Partial Least Squares Modelling, named Partial Least Squares K-Means (PLS-KM), recently proposed by Fordellone and Vichi (2017). In this model, centroids are laying the reduced space of the latent variables, ensuring the optimal partition of the statistical units on the best latent hyperplane. Estimating the measurement relations by the SEM pre-specified model, the latent structure is defined.

2 ESS data

The data from the eighth iteration of the survey for ESS are until now available from 18 of the 24 countries, which undertook fieldwork in 2016. The 18 countries included in this initial release are: Austria, Belgium, Czech Republic, Estonia, Finland, France, Germany, Iceland, Ireland, Israel, Norway, Netherlands, Poland, Russia, Slovenia, Sweden, Switzerland and United Kingdom. The included questions asked in every round since 2002 on topics including crime, democracy and politics, human values, immigration, media consumption, national and ethnic identity, perceived discrimination, religion, social exclusion, social trust/trust in institutions, subjective wellbeing and socio-demographics and public attitudinal data towards welfare, climate change and energy security, personal norms, efficacy and trust and energy preferences. The data must be weighted to adjust for different selection probabilities, for sampling error and non-response bias as well as different selection probabilities. The table 1 shows the topics covered by the survey in the collection of questions, classified into two main parts: a core section and a rotating section. The core module contains items measuring a range of topics of enduring interest to the social sciences as well as the most comprehensive set of socio-structural variables of any cross-national survey. The rotating modules are carried out by multi-national teams of researchers selected to contribute to the design of survey.

Table 1 - *Topics and items of ESS.*

Items	Topics
Core A1-A6	Media use; internet use; social trust
Core B1-B43	Politics, including: political interest, trust, electoral and other forms of participation, party allegiance, socio-political orientations, immigration
Core C1-C44	Subjective wellbeing, social exclusion, crime, religion, perceived discrimination, national and ethnic identity, test questions (sect. I), refugees
Rotating D1-D32	Climate change and energy, including: attitudes, perceptions module and policy preferences
Rotating E1-E42	Welfare, including attitudes towards welfare provision, size of module claimant groups, attitudes towards service delivery and likely future dependence on welfare, vote intention in EU referendum
Core F1-F61	Socio-demographic profile, including household composition, sex, age, marital status, type of area, education and occupation, partner, parents, union membership, income and ancestry
Core Section H	Human values scale
Core Section l	Test questions

Source: www.europeansocialsurvey.org.

The ESS sampling strategy is based on the design and implementation of workable and equivalent sampling plans in all participating countries, following key principles:

- samples must be representative of all persons aged 15 and over (no upper age limit) resident within private households in each country, regardless of their nationality, citizenship or language
- individuals are selected by strict random probability methods at every stage
- sampling frames of individuals, households and addresses may be used
- all countries must aim for a minimum 'effective achieved sample size' of 1,500 or 800 in countries with ESS populations of less than 2 million after discounting for design effects
- quota sampling is not permitted at any stage
- substitution of non-responding households or individuals (whether 'refusals', 'non-contacts' or 'ineligibles') is not permitted at any stage.

In the present paper, we use ESS Multilevel Data resource in order to analyse the ESS-respondents with reference to the context they live in. The resource contains data about:

- individuals (the ESS respondents)
- regions (mainly data collected from EUROSTAT)
- countries (data collected from different sources)

3 Methodology

Given the $n \times J$ data matrix $\mathbf{X}$, the $n \times K$ membership matrix $\mathbf{U}$, the $K \times J$ centroids matrix $\mathbf{C}$, the $J \times P$ loadings matrix $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_{\mathrm{H}}, \boldsymbol{\Lambda}_{\mathrm{L}}]$, and the errors matrices $\mathbf{E}$, $\mathbf{Z}$, $\mathbf{D}$, the Partial Least Squares Structural Equation Modelling *K*-Means approach can be written as follows (Fordellone and Vichi, 2017; Fordellone et al., 2018):

$$\begin{aligned}
&\mathbf{H} = \mathbf{H}\mathbf{B}^{\mathrm{T}} + \boldsymbol{\Xi}\boldsymbol{\Gamma}^{\mathrm{T}} + \mathbf{Z}, \\
&\mathbf{X} = \boldsymbol{\Xi}\boldsymbol{\Lambda}_{\mathrm{H}}^{\mathrm{T}} + \mathbf{H}\boldsymbol{\Lambda}_{\mathrm{L}}^{\mathrm{T}} + \mathbf{E}, \\
&\mathbf{X} = \mathbf{U}\mathbf{C}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}} = \mathbf{U}\mathbf{C}\boldsymbol{\Lambda}_{\mathrm{H}}\boldsymbol{\Lambda}_{\mathrm{H}}^{\mathrm{T}} + \mathbf{U}\mathbf{C}\boldsymbol{\Lambda}_{\mathrm{L}}\boldsymbol{\Lambda}_{\mathrm{L}}^{\mathrm{T}} + \mathbf{D},
\end{aligned} \tag{1}$$

under constraints: *(i)* $\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda} = \mathbf{I}$; and *(ii)* $\mathbf{U} \in \{0,1\}$, $\mathbf{U}\mathbf{1}_K = \mathbf{1}_n$. Where, H is the n×L matrix of the endogenous LVs with generic element $\eta_{i.l}$, $\boldsymbol{\Xi}$ be the $n \times H$ matrix of the exogenous LVs with generic element $\xi_{i.h}$, $\mathbf{B}$ is the $L \times L$ matrix of the path coefficients $\beta_{l.l}$ associated to the endogenous latent variables, $\boldsymbol{\Gamma}$ is the $L \times H$ matrix of the path coefficients $\gamma_{l.h}$ associated to the exogenous latent variables, $\boldsymbol{\Lambda}_{\mathrm{H}}$ is the $J \times H$ loadings matrix of the exogenous latent constructs with generic element $\lambda_{i.h}$, and $\boldsymbol{\Lambda}_{\mathrm{L}}$is the $J \times L$ loadings matrix of the endogenous latent constructs with generic element $\lambda_{i.l}$. Thus, the PLS-SEM-KM model includes the SEM estimated via Partial Least Squares (PLS) and the clustering equations. The simultaneous estimation of the three sets of equations will produce the estimation of the pre-specified SEM describing relations among variables and the corresponding best partitioning of units

There is a relevant aspect to considerate in the application of PLS-SEM-KM procedure: when we applying PLS-SEM-KM, the number of groups is unknown and the identification of an appropriate number of K clusters is not straightforward. Then, often you need to rely on some statistical criterion. In particular, the PLS-SEM-KM algorithm includes the choice of the number of clusters K classes according the *gap method* criterion (Fordellone and Vichi, 2017).

References

ENTZINGER, H. 2000. The Dynamics of Integration Policies: A Multidimensional Model. Challenging Immigration and Ethnic Relations Politics. *R. Koopmans and P. Statham, Oxford: University Press*.

FORDELLONE, M., VICHI, M. 2017. Partial Least Squares Modelling and simultaneous clustering. *Cladag 2017 Book of Short Papers. Universitas Studiorum*.

FORDELLONE, M., TOMASELLI V., AND VICHI M. 2018. From Tandem to Simultaneous Dimensionality Reduction And Clustering Of Tourism Data. *Rivista Italiana di Economia Demografia e Statistica 72.1*.

TENENHAUS, M., VINZI, E.V., CHATELIN, Y.M., LAURO, N.C. 2005. PLS path modeling. *Computational Statistics & Data Analysis. Vol. 48 No. 1, pp. 159–205*.

CLADAG 2019
Cassino (ITALY)
11-13 September, 2019

Società Italiana di Statistica

UNICAS 1979>2019
40 YEARS FROM FOUNDATION
UNIVERSITÀ DI CASSINO E DEL LAZIO MERIDIONALE

The CLAssification and Data Analysis Group of the Italian Statistical Society (SIS) 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.

CLADAG is a member of the International Federation of Classification Societies (IFCS).

Among its activities, CLADAG organizes a biennial international 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.

Founded in 1985, the IFCS is a federation of national, regional, and linguistically-based classification societies. It is a non-profit, nonpolitical scientific organization, whose aims are to further classification research.