

SIS 2017
Statistics and Data Science:
new challenges, new generations

28–30 June 2017
Florence (Italy)

Proceedings of the Conference
of the Italian Statistical Society

edited by
Alessandra Petrucci
Rosanna Verde

FIRENZE UNIVERSITY PRESS
2017

SIS 2017. Statistics and Data Science: new challenges, new generations : 28-30 June 2017 Florence (Italy) : proceedings of the Conference of the Italian Statistical Society / edited by Alessandra Petrucci, Rosanna Verde. – Firenze : Firenze University Press, 2017.

(Proceedings e report ; 114)

<http://digital.casalini.it/9788864535210>

ISBN 978-88-6453-521-0 (online)

Peer Review Process

All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Committees of the individual series. The works published in the FUP catalogue are evaluated and approved by the Editorial Board of the publishing house. For a more detailed description of the refereeing process we refer to the official documents published on the website and in the online catalogue of the FUP (www.fupress.com).

Firenze University Press Editorial Board

A. Dolfi (Editor-in-Chief), M. Boddi, A. Bucelli, R. Casalbuoni, M. Garzaniti, M.C. Grisolia, P. Guarnieri, R. Lanfredini, A. Lenzi, P. Lo Nostro, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, G. Nigro, A. Perulli, M.C. Torricelli.

This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0: <https://creativecommons.org/licenses/by/4.0/legalcode>)

CC 2017 Firenze University Press
Università degli Studi di Firenze
Firenze University Press
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

Index

Preface	XXV
Alexander Agapitov, Irina Lackman, Zoya Maksimenko <i>Determination of basis risk multiplier of a borrower default using survival analysis</i>	1
Tommaso Agasisti, Alex J. Bowers, Mara Soncin <i>School principals' leadership styles and students achievement: empirical results from a three-step Latent Class Analysis</i>	7
Tommaso Agasisti, Sergio Longobardi, Felice Russo <i>Poverty measures to analyse the educational inequality in the OECD Countries</i>	17
Mohamed-Salem Ahmed, Laurence Broze, Sophie Dabo-Niang, Zied Gharbi <i>Quasi-Maximum Likelihood Estimators For Functional Spatial Autoregressive Models</i>	23
Giacomo Aletti, Alessandra Micheletti <i>A clustering algorithm for multivariate big data with correlated components</i>	31
Emanuele Aliverti <i>A Bayesian semiparametric model for terrorist networks</i>	37

- Giorgio Alleva
Emerging challenges in official statistics: new sources, methods and skills 43
- Rémi André, Xavier Luciani and Eric Moreau
A fast algorithm for the canonical polyadic decomposition of large tensors 45
- Maria Simona Andreano, Roberto Benedetti, Paolo Postiglione, Giovanni Savio
On the use of Google Trend data as covariates in nowcasting: Sampling and modeling issues 53
- Francesco Andreoli, Mauro Mussini
A spatial decomposition of the change in urban poverty concentration 59
- Margaret Antonicelli, Vito Flavio Covella
How green advertising can impact on gender different approach towards sustainability 65
- Rosa Arboretti, Eleonora Carrozzo, Luigi Salmaso
Stratified data: a permutation approach for hypotheses testing 71
- Marika Arena, Anna Calissano, Simone Vantini
Crowd and Minorities: Is it possible to listen to both? Monitoring Rare Sentiment and Opinion Categories about Expo Milano 2015 79
- Maria Felice Arezzo, Giuseppina Guagnano
Using administrative data for statistical modeling: an application to tax evasion 83
- Monica Bailot, Rina Camporese, Silvia Da Valle, Sara Letardi, Susi Osti
Are Numbers too Large for Kids? Possible Answers in Probable Stories 89

Index	IX
Simona Balbi, Michelangelo Misuraca, Germana Scepti <i>A polarity-based strategy for ranking social media reviews</i>	95
A. Balzanella, S.A. Gattone, T. Di Battista, E. Romano, R. Verde <i>Monitoring the spatial correlation among functional data streams through Moran's Index</i>	103
Oumayma Banouar, Saïd Raghay <i>User query enrichment for personalized access to data through ontologies using matrix completion method</i>	109
Giulia Barbati, Francesca Ieva, Francesca Gasperoni, Annamaria Iorio, Gianfranco Sinagra, Andrea Di Lenarda <i>The Trieste Observatory of cardiovascular disease: an experience of administrative and clinical data integration at a regional level</i>	115
Francesco Bartolucci, Stefano Peluso, Antonietta Mira <i>Marginal modeling of multilateral relational events</i>	123
Francesca Bassi, Leonardo Grilli, Omar Paccagnella, Carla Rampichini, Roberta Varriale <i>New Insights on Students Evaluation of Teaching in Italy</i>	129
Mauro Bernardi, Marco Bottone, Lea Petrella <i>Bayesian Quantile Regression using the Skew Exponential Power Distribution</i>	135
Mauro Bernardi <i>Bayesian Factor-Augmented Dynamic Quantile Vector Autoregression</i>	141

- Bruno Bertaccini, Giulia Biagi, Antonio Giusti, Laura Grassini
Does data structure reflect monuments structure? Symbolic data analysis on Florence Brunelleschi Dome
149
- Gaia Bertarelli and Franca Crippa, Fulvia Mecatti
A latent markov model approach for measuring national gender inequality
157
- Agne Bikauskaite, Dario Buono
Eurostat's methodological network: Skills mapping for a collaborative statistical office
161
- Francesco C. Billari, Emilio Zagheni
Big Data and Population Processes: A Revolution?
167
- Monica Billio, Roberto Casarin, Matteo Iacopini
Bayesian Tensor Regression models
179
- Monica Billio, Roberto Casarin, Luca Rossini
Bayesian nonparametric sparse Vector Autoregressive models
187
- Chiara Bocci, Daniele Fadda, Lorenzo Gabrielli, Mirco Nanni, Leonardo Piccini
Using GPS Data to Understand Urban Mobility Patterns: An Application to the Florence Metropolitan Area
193
- Michele Boreale, Fabio Corradi
Relative privacy risks and learning from anonymized data
199
- Giacomo Bormetti, Roberto Casarin, Fulvio Corsi, Giulia Livieri
A stochastic volatility framework with analytical filtering
205

Index	XI
Alessandro Brunetti, Stefania Fatello, Federico Polidoro <i>Estimating Italian inflation using scanner data: results and perspectives</i>	211
Guénael Cabanes, Younès Bennani, Rosanna Verde, Antonio Irpino <i>Clustering of histogram data : a topological learning approach</i>	219
Renza Campagni, Lorenzo Gabrielli, Fosca Giannotti, Riccardo Guidotti, Filomena Maggino, Dino Pedreschi <i>Measuring Wellbeing by extracting Social Indicators from Big Data</i>	227
Maria Gabriella Campolo, Antonino Di Pino <i>Assessing Selectivity in the Estimation of the Causal Effects of Retirement on the Labour Division in the Italian Couples</i>	235
Stefania Capecchi, Rosaria Simone <i>Composite indicators for ordinal data: the impact of uncertainty</i>	241
Stefania Capecchi, Domenico Piccolo <i>The distribution of Net Promoter Score in socio-economic surveys</i>	247
Massimiliano Caporin, Francesco Poli <i>News, Volatility and Price Jumps</i>	253
Carmela Cappelli, Rosaria Simone, Francesca di Iorio <i>Growing happiness: a model-based tree</i>	261
Paolo Emilio Cardone <i>Inequalities in access to job-related learning among workers in Italy: evidence from Adult Education Survey (AES)</i>	267

- Alessandro Casa, Giovanna Menardi
Signal detection in high energy physics via a semisupervised nonparametric approach
273
- Claudio Ceccarelli, Silvia Montagna, Francesca Petrarca
Employment study methodologies of Italian graduates through the data linkage of administrative archives and sample surveys
279
- Ikram Chairi, Amina El Gonnouni, Sarah Zouinina, Abdelouahid Lyhyaoui
Prediction of Firm's Creditworthiness Risk using Feature Selection and Support Vector Machine
285
- Sana Chakri, Said Raghay, Salah El Hadaj
Contribution of extracting meaningful patterns from semantic trajectories
293
- Chieppa A., Ferrara R., Gallo G., Tomeo V.
Towards The Register-Based Statistical System: A New Valuable Source for Population Studies
301
- Shirley Coleman
Consulting, knowledge transfer and impact case studies of statistics in practice
305
- Michele Costa
The evaluation of the inequality between population subgroups
313
- Michele Costola
Bayesian Non-Negative l_1 -Regularised Regression
319
- Lisa Crosato, Caterina Liberati, Paolo Mariani, Biancamaria Zavarella
Industrial Production Index and the Web: an explorative cointegration analysis
327

Index	XIII
Francesca Romana Crucinio, Roberto Fontana <i>Comparison of conditional tests on Poisson data</i>	333
Riccardo D'Alberto, Meri Raggi <i>Non-parametric micro Statistical Matching techniques: some developments</i>	339
Stefano De Cantis, Mauro Ferrante, Anna Maria Parroco <i>Measuring tourism from demand side</i>	345
Lucio De Capitani, Daniele De Martini <i>Optimal Ethical Balance for Phase III Trials Planning</i>	351
Claudia De Vitiis, Alessio Guandalini, Francesca Inglese, Marco D. Terribili <i>Sampling schemes using scanner data for the consumer price index</i>	357
Ermelinda Della Valle, Elena Scardovi, Andrea Iacobucci, Edoardo Tignone <i>Interactive machine learning prediction for budget allocation in digital marketing scenarios</i>	365
Marco Di Marzio, Stefania Fensore, Agnese Panzera, Charles C. Taylor <i>Nonparametric classification for directional data</i>	371
Edwin Diday <i>Introduction to Symbolic Data Analysis and application to post clustering for comparing and improving clustering methods by the Symbolic Data Table that they induce</i>	379
Carlo Drago <i>Identifying Meta Communities on Large Networks</i>	387

- Neska El Haouij, Jean-Michel Poggi, Raja Ghozi, Sylvie Sevestre Ghalila, Mériem Jaidane
Random Forest-Based Approach for Physiological Functional Variable Selection for Drivers Stress Level Classification
393
- Silvia Facchinetti, Silvia A. Osmetti
A risk index to evaluate the criticality of a product defectiveness
399
- Federico Ferraccioli, Livio Finos
Exponential family graphical models and penalizations
405
- Mauro Ferrante, Giovanna Fantaci, Anna Maria Parroco, Anna Maria Milito, Salvatore Scondotto
Key-indicators for maternity hospitals and newborn readmission in Sicily
411
- Ferretti Camilla, Ganugi Piero, Zammori Francesco
Change of Variables theorem to fit Bimodal Distributions
417
- Francesco Finazzi, Lucia Paci
Space-time clustering for identifying population patterns from smartphone data
423
- Annunziata Fiore, Antonella Simone, Antonino Virgillito
IT Solutions for Analyzing Large-Scale Statistical Datasets: Scanner Data for CPI
429
- Michael Fop, Thomas Brendan Murphy, Luca Scrucca
Model-based Clustering with Sparse Covariance Matrices
437
- Maria Franco-Villoria, Marian Scott
Quantile Regression for Functional Data
441

Index	XV
Gallo M., Simonacci V., Di Palma M.A. <i>Three-way compositional data: a multi-stage trilinear decomposition algorithm</i>	445
Francesca Gasperoni, Francesca Ieva, Anna Maria Paganoni, Chris Jackson, Linda Sharples <i>Nonparametric shared frailty model for classification of survival data</i>	451
Stefano A. Gattone, Angela De Sanctis <i>Clustering landmark-based shapes using Information Geometry tools</i>	457
Alan E. Gelfand, Shinichiro Shirota <i>Space and circular time log Gaussian Cox processes with application to crime event data</i>	461
Abdelghani Ghazdali <i>Blind source separation</i>	469
Massimiliano Giacalone, Antonio Ruoto, Davide Liga, Maria Pilato, Vito Santarangelo <i>An innovative approach for Opinion Mining : the Plutchick analysis</i>	479
Massimiliano Giacalone, Demetrio Panarello <i>A G.E.D. method for market risk evaluation using a modified Gaussian Copula</i>	485
Chiara Gigliarano, Francesco Maria Chelli <i>Labour market dynamics and recent economic changes: the case of Italy</i>	491
Giuseppe Giordano, Giancarlo Ragozini, Maria Prosperina Vitale <i>On the use of DISTATIS to handle multiplex networks</i>	499

- Michela Gnaldi, Silvia Bacci, Samuel Greiff, Thiemo Kunze
Profiles of students on account of complex problem solving (CPS) strategies exploited via log-data
505
- Michela Gnaldi, Simone Del Sarto
Characterising Italian municipalities according to the annual report of the prevention-of-corruption supervisor: a Latent Class approach
513
- Silvia Golia
A proposal of a discretization method applicable to Rasch measures
519
- Anna Gottard
Tree-based Non-linear Graphical Models
525
- Sara Hbali, Youssef Hbali, Mohamed Sadgal, Abdelaziz El Fazziki
Sentiment Analysis for micro-blogging using LSTM Recurrent Neural Networks
531
- Stefano Maria Iacus, Giuseppe Porro, Silvia Salini, Elena Siletti
How to Exploit Big Data from Social Networks: a Subjective Well-being Indicator via Twitter
537
- Francesca Ieva
Network Analysis of Comorbidity Patterns in Heart Failure Patients using Administrative Data
543
- Antonio Irpino, Francisco de A.T. De Carvalho, Rosanna Verde
Automatic variable and components weighting systems for Fuzzy cmeans of distributional data
549
- Michael Jauch, Paolo Giordani, David Dunson
A Bayesian oblique factor model with extension to tensor data
553

Index	XVII
Johan Koskinen, Chiara Broccatelli, Peng Wang, Garry Robins <i>Statistical analysis for partially observed multilayered networks</i>	561
Francesco Lagona <i>Copula-based segmentation of environmental time series with linear and circular components</i>	569
Alessandro Lanteri, Mauro Maggioni <i>A Multiscale Approach to Manifold Estimation</i>	575
Tiziana Laureti, Carlo Ferrante, Barbara Dramis <i>Using scanner and CPI data to estimate Italian sub-national PPPs</i>	581
Antonio Lepore <i>Graphical approximation of Best Linear Unbiased Estimators for Extreme Value Distribution Parameters</i>	589
Antonio Lepore, Biagio Palumbo, Christian Capezza <i>Monitoring ship performance via multi-way partial least-squares analysis of functional data</i>	595
Caterina Liberati, Lisa Crosato, Paolo Mariani, Biancamaria Zavanella <i>Dynamic profiling of banking customers: a pseudo-panel study</i>	601
Giovanni L. Lo Magno, Mauro Ferrante, Stefano De Cantis <i>A comparison between seasonality indices deployed in evaluating unimodal and bimodal patterns</i>	607
Rosaria Lombardo, Eric J Beh <i>Three-way Correspondence Analysis for Ordinal-Nominal Variables</i>	613

- Monia Lupparelli, Alessandra Mattei
Log-mean linear models for causal inference
621
- Badiaa Lyoussi, Zineb Selihi, Mohamed Berraho, Karima El Rhazi, Youness El Achhab, Adiba El Marrakchi, Chakib Nejjari
Research on the Risk Factors accountable for the occurrence of degenerative complications of type 2 diabetes in Morocco: a prospective study
627
- Valentina Mameli, Debora Slanzi, Irene Poli
Bootstrap group penalty for high-dimensional regression models
633
- Stefano Marchetti, Monica Pratesi, Caterina Giusti
Improving small area estimates of households' share of food consumption expenditure in Italy by means of Twitter data
639
- Paolo Mariani, Andrea Marletta, Mariangela Zenga
Gross Annual Salary of a new graduate: is it a question of profile?
647
- Maria Francesca Marino, Marco Alfò
Dynamic random coefficient based drop-out models for longitudinal responses
653
- Antonello Maruotti, Jan Bulla
Hidden Markov models: dimensionality reduction, atypical observations and algorithms
659
- Chiara Masci, Geraint Johnes, Tommaso Agasisti
A flexible analysis of PISA 2015 data across countries, by means of multilevel trees and boosting
667

Index	XIX
Lucio Masserini, Matilde Bini <i>Impact of the 2008 and 2012 financial crises on the unemployment rate in Italy: an interrupted time series approach</i>	673
Angelo Mazza, Antonio Punzo, Salvatore Ingrassia <i>An R Package for Cluster-Weighted Models</i>	681
Antonino Mazzeo, Flora Amato <i>Methods and applications for the treatment of Big Data in strategic fields</i>	687
Letizia Mencarini, Viviana Patti, Mirko Lai, Emilio Sulis <i>Happy parents' tweets</i>	693
Rodolfo Metulini, Marica Manisera, Paola Zuccolotto <i>Space-Time Analysis of Movements in Basketball using Sensor Data</i>	701
Giorgio E. Montanari, Marco Doretto, Francesco Bartolucci <i>An ordinal Latent Markov model for the evaluation of health care services</i>	707
Isabella Morlini, Maristella Scorza <i>New fuzzy composite indicators for dyslexia</i>	713
Fionn Murtagh <i>Big Textual Data: Lessons and Challenges for Statistics</i>	719
Gaetano Musella, Gennaro Punzo <i>Workers' skills and wage inequality: A time-space comparison across European Mediterranean countries</i>	731

Marta Nai Ruscone <i>Exploratory factor analysis of ordinal variables: a copula approach</i>	737
Fausta Ongaro, Silvana Salvini <i>IPUMS Data for describing family and household structures in the world</i>	743
Tullia Padellini, Pierpaolo Brutti <i>Topological Summaries for Time-Varying Data</i>	747
Sally Paganin <i>Modeling of Complex Network Data for Targeted Marketing</i>	753
Francesco Palumbo, Giancarlo Ragozini <i>Statistical categorization through archetypal analysis</i>	759
Michela Eugenia Pasetto, Umberto Noè, Alessandra Luati, Dirk Husmeier <i>Inference with the Unscented Kalman Filter and optimization of sigma points</i>	767
Xanthi Pedeli, Cristiano Varin <i>Pairwise Likelihood Inference for Parameter-Driven Models</i>	773
Felicia Pelagalli, Francesca Greco, Enrico De Santis <i>Social emotional data analysis. The map of Europe</i>	779
Alessia Pini, Lorenzo Spreafico, Simone Vantini, Alessandro Vietti <i>Differential Interval-Wise Testing for the Inferential Analysis of Tongue Profiles</i>	785
Alessia Pini, Aymeric Stamm, Simone Vantini <i>Hotelling meets Hilbert: inference on the mean in functional Hilbert spaces</i>	791

Index	XXI
Silvia Poletini, Serena Arima <i>Accounting for measurement error in small area models: a study on generosity</i>	795
Gennaro Punzo, Mariateresa Ciommi <i>Structural changes in the employment composition and wage inequality: A comparison across European countries</i>	801
Walter J. Radermacher <i>Official Statistics 4.0 – learning from history for the challenges of the future</i>	809
Fabio Rapallo <i>Comparison of contingency tables under quasi-symmetry</i>	821
Valentina Raponi, Cesare Robotti, Paolo Zaffaroni <i>Testing Beta-Pricing Models Using Large Cross-Sections</i>	827
Marco Seabra dos Reis, Biagio Palumbo, Antonio Lepore, Ricardo Rendall, Christian Capezza <i>On the use of predictive methods for ship fuel consumption analysis from massive on-board operational data</i>	833
Alessandra Righi, Mauro Mario Gentile <i>Twitter as a Statistical Data Source: an Attempt of Profiling Italian Users Background Characteristics</i>	841
Paolo Righi, Giulio Barcaroli, Natalia Golini <i>Quality issues when using Big Data in Official Statistics</i>	847
Emilia Rocco <i>Indicators for the representativeness of survey response as well as convenience samples</i>	855

- Emilia Rocco, Bruno Bertaccini, Giulia Biagi, Andrea Giommi
A sampling design for the evaluation of earthquakes vulnerability of the residential buildings in Florence
861
- Elvira Romano, Jorge Mateu
A local regression technique for spatially dependent functional data: an heteroskedastic GWR model
867
- Eduardo Rossi, Paolo Santucci de Magistris
Models for jumps in trading volume
873
- Renata Rotondi, Elisa Varini
On a failure process driven by a self-correcting model in seismic hazard assessment
879
- M. Ruggieri, F. Di Salvo and A. Plaia
Functional principal component analysis of quantile curves
887
- Massimiliano Russo
Detecting group differences in multivariate categorical data
893
- Michele Scagliarini
A Sequential Test for the C_{pk} Index
899
- Steven L. Scott
Industrial Applications of Bayesian Structural Time Series
905
- Catia Scricciolo
Asymptotically Efficient Estimation in Measurement Error Models
913

Index	XXIII
Angela Serra, Pietro Coretto, Roberto Tagliaferri <i>On the noisy high-dimensional gene expression data analysis</i>	919
Mirko Signorelli <i>Variable selection for (realistic) stochastic blockmodels</i>	927
Marianna Siino, Francisco J. Rodriguez-Cortés, Jorge Mateu, Giada Adelfio <i>Detection of spatio-temporal local structure on seismic data</i>	935
A. Sottosanti, D. Bastieri, A. R. Brazzale <i>Bayesian Mixture Models for the Detection of High-Energy Astronomical Sources</i>	943
Federico Mattia Stefanini <i>Causal analysis of Cell Transformation Assays</i>	949
Paola Stolfi, Mauro Bernardi, Lea Petrella <i>Estimation and Inference of SkewStable distributions using the Multivariate Method of Simulated Quantiles</i>	955
Paola Stolfi, Mauro Bernardi, Lea Petrella <i>Sparse Indirect Inference</i>	961
Peter Struijs, Anke Consten, Piet Daas, Marc Debusschere, Maiki Ilves, Boro Nikic, Anna Nowicka, David Salgado, Monica Scannapieco, Nigel Swier <i>The ESSnet Big Data: Experimental Results</i>	969
Jérémie Sublime <i>Smart view selection in multi-view clustering</i>	977

- Emilio Sulis
Social Sensing and Official Statistics: call data records and social media sentiment analysis
985
- Matilde Trevisani, Arjuna Tuzzi
Knowledge mapping by a functional data analysis of scientific articles databases
993
- Amalia Vanacore, Maria Sole Pellegrino
Characterizing the extent of rater agreement via a non-parametric benchmarking procedure
999
- Maarten Vanhoof, Stephanie Combes, Marie-Pierre de Bellefon
Mining Mobile Phone Data to Detect Urban Areas
1005
- Viktoriya Voytsekhovska, Olivier Butzbach
Statistical methods in assessing the equality of income distribution, case study of Poland
1013
- Ernst C. Wit
Network inference in Genomics
1019
- Dilek Yildiz, Jo Munson, Agnese Vitali, Ramine Tinati, Jennifer Holland
Using Twitter data for Population Estimates
1025
- Marco Seabra dos Rei
Structured Approaches for High-Dimensional Predictive Modeling
1033

Using administrative data for statistical modeling: an application to tax evasion

L'uso di dati amministrativi per la modellizzazione statistica: un'applicazione all'evasione contributiva

Maria Felice Arezzo and Giuseppina Guagnano

Abstract Administrative data, gathered by public authorities with a general aim of control, are very precious sources of information because they allow to study phenomena that would remain otherwise unknown. On the other side, administrative data strictly contain the information they were collected for, and to be used for statistical purposes they need to be integrated. This work shows the potentials of the integration of three data sets for statistical modeling: the audits carried out in Italy in 2005 by the National Institute of Social Security on building and construction companies, the ASIA archive of Istat and the "Studi di Settore" of the Italian Revenue Agency.

Abstract *I dati amministrativi, raccolti dalle istituzioni pubbliche per scopi generalmente di controllo, sono fonti informative estremamente preziose in quanto permettono spesso di studiare fenomeni che in altro modo non potrebbero essere conosciuti. D'altro canto, proprio perchè rispondono a finalità specifiche, le indagini amministrative non contengono informazioni aggiuntive rispetto a quelle per le quali sono state pensate. Il lavoro illustra le potenzialità dell'integrazione di tre basi dati da fonte differente: le ispezioni INPS, l'archivio ASIA dell'Istat e gli Studi di settore dell'Agenzia delle entrate. La sperimentazione è stata condotta sulle imprese che operano nel settore delle costruzioni.*

Key words: Administrative data, Sample selection, Response-based sampling

Maria Felice Arezzo

Sapienza University of Rome, Address, e-mail: mariafelice.arezzo@uniroma1.it

Giuseppina Guagnano

Sapienza University of Rome, Address e-mail: giuseppina.guagnano@uniroma1.it

1 Introduction

Administrative data are archives of great interest as they often contain information available only to public authorities responsible for the control of some phenomena. Almost always, though, these files do not contain information other than those for which they were collected (a typical example are the socio-economic characteristics of the individuals), as the purpose underlying their gathering is not statistical modeling. For this very same reason, administrative data require, on the one side, a throughout pretreatment and validation process and, on the other, the development of statistical methodologies that allow for the drawing of valid inferences.

The purpose of our work is to draw the entire “production chain”: a) the creation of a dataset with all relevant variables, b) the evaluation of the dataset quality, c) the development of a statistical method suitable for the data at stake.

The case study is on the detection of the firms which evade worker contributions because they employ off-the-book workers (i.e. employee who are completely unknown to fiscal authorities)

2 Creation of the data set

Our starting point is an administrative dataset on the audits carried out in Italy in 2005 by the National Institute of Social Security (INPS henceforth) on building and construction companies (NACE section: F). It amounts to a total of 31,658 inspections on 28,731 firms. The global amount of firms operating in the building industry in Italy in the same year was $N = 595,226$. Audits data allow to observe the compliant/non-compliant behavior.

Following the idea that the risk of a non-compliant behavior can be predicted by the economic characteristics of the firm, we integrated the information of audits with two other sources of data. The first is the ASIA archive owned by the National Institute of Statistics (ISTAT). It contains data on the legal structure, turnover and number of employee and is a high quality source of data as the information are validated through a very careful process. The second, owned by the Italian Revenue Agency, is the so called ‘Studi di Settore’ (SS in the following) archive. It contains an exhaustive list of information on corporate organization, firm structure, management and governance.

The three data sets were merged using VAT numbers and/or tax codes. Surprisingly the match rate was only 51% meaning that the number of firms in the merged archive is 14,651.

The original variables were used to build economic indicators which can be grouped in the following different firm’s facets: a) 9 indicators for economic dimension, b) 13 for organization, c) 6 for structure, d) 6 for management, e) 11 for performance f) 38 for labor productivity and profitability g) 3 for contracts award mode h) 7 variables for location and type. The final dataset had 93 independent variables observed on 14,651 building companies with a match rate of 51%. The

Table 1 Datasets characteristics

Data Owner	Content	Individual	Dimension
INPS	Inspections outputs (2005)	Inspection	31,658 inspections on 28,731 firms
Revenue Agency	Studi di settore (2005).Models: TG69U, TG75U (SG75U),TG50U (SG50U and SG71U), TG70U	Firm	Universe of firms with at most 5 million euros of income
ISTAT	Asia Archives (2005)	Firm	Universe of firms

variable to be predicted is named Y and it takes value 1 if in a firm there is at least one off-the-book worker and 0 otherwise. In the following we will refer to the final dataset as the integrated db because it gathers and integrate information from different sources.

3 The assessment of the integrated dataset

As we said, the matching rate was 51% which means that we had information on the features of interest for (roughly) half of the firms in original INPS database. We studied inspection coverage and the risk of non complying for different turnover class and corporate designation typologies and over the territory. The idea was to verify if a whole group of firms (for example all the companies in a geographical region) was lost because of the merging process.

We checked for: Regions (20 levels), Number of employee (9 classes), Legal structure (5 levels), Turnover (11 classes); we then made sure that during the matching procedure, no whole groups of individuals were lost.

4 The Model

Under a statistical point of view, there are two main methodological issues arising from the type of data we use. The first is the non-randomness of the inspections and the second is that the fraction of inspected firms in the population is low.

SELECTION BIAS IN THE SAMPLE OF INSPECTED FIRM. To detect undeclared work, an inspector audits firms. Inspected firms are not randomly chosen; they are chosen because the inspector thinks that there are some off-the-book workers and s/he has strong incentives to target the “right” firms (i.e. the irregular ones). We can think of the decision to inspect a firm as a rational process in which the inspection is made if the utility to inspect, U^A , (i.e. find undeclared workers and get a benefit) is higher than the utility of non-inspect, $U^{\bar{A}}$. Moreover we can observe the status of

the i -th firm (regular or not) only if it has been inspected, otherwise a censoring process intervenes. It is obvious that there is a strong selection bias in the sample of inspected firms.

As it is well known, [3] proposed a useful framework for handling estimation when the sample is subject to a selection mechanism. In the original framework, the outcome variable is continuous and can be explained by a linear regression model (called *output equation*), with a normal random component; in addition to the output equation, a *selection equation* describes the selection rule by means of a binary choice model (probit).

In our framework the output equation defines the compliance decision, so the dependent variable is binary, and the selection equation refers to the decision of inspecting a firm. Just as the inspection decision, the evasion is based on a rational process and it happens if the utility of evading, $U_i^{\bar{C}}$, is greater than the utility of complying U_i^C . The corresponding econometric model, in its general form, is:

$$Y_i^* = U_i^{\bar{C}} - U_i^C = \mathbf{X}_{1i}\beta + \varepsilon_{1i} \quad (1a)$$

$$A_i^* = U_i^A - U_i^{\bar{A}} = \mathbf{X}_{2i}\theta + \varepsilon_{2i} \quad (1b)$$

where $\mathbf{X}_i = (\mathbf{X}_{1i}, \mathbf{X}_{2i})$ is a vector of exogenous variables (namely, \mathbf{X}_{1i} for Y_i and \mathbf{X}_{2i} for A_i), containing all the relevant covariates.

Since we cannot observe directly the utilities (neither those determining compliance, nor those governing the decision to inspect), we assume that if in equation (1a) $Y_i^* > 0$, the firm does not comply, otherwise it does. Let's define a dummy variable Y_i which we can observe and that denotes the alternative selected:

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Similarly, we can define an observable dummy variable A_i for the inspections, such that:

$$A_i = \begin{cases} 1 & \text{if } A_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The p.d.f. of Y_i and A_i is Bernoulli with probability of success respectively equal to $\gamma\pi$ and ${}_A\pi$ and depending on $\mathbf{X}_{1i}\beta$ and on $\mathbf{X}_{2i}\theta$. A selection bias exists if $\text{corr}(\varepsilon_1, \varepsilon_2) = \rho$ is not null.

As it is known (see for example [2]), the likelihood function for the Heckman's selection model is:

$$L(\eta) = \prod_{i=1}^n \left[1 - {}_A\pi(\mathbf{X}_i) \right]^{1-A_i} \cdot \left[f(Y_i|A_i = 1) \cdot {}_A\pi(\mathbf{X}_i) \right]^{A_i} \quad (4)$$

where $\eta = (\beta, \theta, \rho)$ is the vector of parameters to be estimated.

THE CASE-CONTROL SETTING. In this sampling design [4], also known as response-based, samples of fixed size are randomly chosen from the two strata identified by the dependent variable A . In particular n_A units are drawn at random from the N_A cases and $n_{\bar{A}}$ from the $N_{\bar{A}}$ controls.

The likelihood function is the product of the two stratum-specific likelihoods and depends on the probability that the individual is in the sample, and on the joint density of the covariates:

$$\prod_{i=1}^{n_A} Pr(\mathbf{X}_i | A_i = 1, S_i = 1) \cdot \prod_{i=1}^{n_{\bar{A}}} Pr(\mathbf{X}_i | A_i = 0, S_i = 1). \quad (5)$$

The c-c design is particularly suited in our study because the probability that a firm is inspected is very low and therefore it is much more convenient to directly sample from the two strata (inspected/non-inspected).

A BINARY CHOICE MODEL WITH SAMPLE SELECTION AND CASE-CONTROL SAMPLING SCHEME. In the following we provide the likelihood function under the framework of interest, i.e. a sample selection mechanism with a severe censoring process. The interested reader can find the full proof and the simulation results in [1].

We make the following very general and non restrictive assumptions:

1. we have a set of fully informative and exogenous covariates $\mathbf{X}_i = (\mathbf{X}_{1i}, \mathbf{X}_{2i})$;
2. conditional on the covariates, the probability that an observation is uncensored doesn't depend on its value, i.e. $P(A_i = 1 | S_i = 1, \mathbf{X}_i, Y_i) = P(A_i = 1 | S_i = 1, \mathbf{X}_i)$;
3. the set of covariates \mathbf{X}_{1i} , specific for Y_i , and the set \mathbf{X}_{2i} , specific for A_i , may have common elements but they cannot fully overlap;
4. the probability of being in the sample does not depends neither on the covariates \mathbf{X}_i nor on Y_i . More precisely, letting S_i be a binary variable which takes value 1 if the i -th individual is in the sample and 0 otherwise, it is true that $P(S_i = 1 | \mathbf{X}_i, Y_i, A_i = a_i) = P(S_i = 1 | A_i = a_i)$.

Assumption (1) means that it does not exist correlation between the covariates and the residual terms in equations (1a) and (1b). Assumption 2 is justified because, as the covariates are informative, all the information brought by Y_i is contained in \mathbf{X}_i . Assumption (3) is necessary for parameters identification (exclusion conditions). Assumption (4) is typical in the response-based sampling framework and no further explanation is required.

Under the conditions stated, the likelihood function for a binary choice model with sample selection under a response-based sampling is:

$$L(\eta) = \prod_{i=1}^n f(\mathbf{X}_i | S_i = 1) \left\{ (1 - {}_A\pi(\mathbf{X}_{2i})) \cdot \frac{N}{N_{\bar{A}}} \right\}^{1-A_i} \quad (6)$$

$$\cdot \left\{ \left[{}_Y\pi(\mathbf{X}_{1i}) \cdot \frac{N}{N_A} \frac{n_A}{n_{1A}} \right]^{y_i} \cdot \left[(1 - {}_Y\pi(\mathbf{X}_{1i})) \cdot \frac{N}{N_A} \frac{n_A}{n_{0A}} \right]^{1-y_i} \cdot {}_A\pi(\mathbf{X}_{2i}) \right\}^{A_i}$$

where ${}_A\pi(\mathbf{X}_{2i})$ is the probability that an observation is uncensored and ${}_Y\pi(\mathbf{X}_{1i})$ is the probability of observing $Y = 1$ given that the observation is uncensored; as already said, n_A is the number of units sampled from the N_A uncensored observations and $n_{\bar{A}}$ is the number of units sampled from the $N_{\bar{A}}$ censored observations; n_{yA} is the amount of units in the sample having $Y = y$, with $y = 0, 1$.

It's easy to understand that the likelihood (6) is a weighted version of (4), and the weights simply take into account the sampling design. Note also that in the maximization process the term $f(\mathbf{X}_i | S_i = 1)$ is non influential, as it does not contain any information on the vector of parameters η , and that in our estimator the only quantities to be known at the population level are N_A and N .

References

1. Arezzo, M.F., Guagnano, G. Response-based sampling for binary choice models with sample selection. Working Paper 149, Department Memotef - Sapienza University of Rome (2017).
2. Cameron, A.C., Trivedi, P.K.: Microeconometrics: Methods and Applications. Cambridge University Press, New York (2005)
3. Heckman, J.J.: Sample selection bias as a specification error. *Econometrica*, 47(1): 153-162 (1979).
4. Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression. John Wiley & Sons (2013)