Data Science & Social Research 2019

Book of Abstracts

Second international conference on data science and social research

Editor: Paolo Mariani

Data Science & Social Research 2019

Book of Abstracts

Second international conference on data science and social research

4 February 2019 - University of Milano - Bicocca

5 February 2019 - Università IULM

Editor: Paolo Mariani

Con il Patrocinio di

Regione Lombardia

PATROCINIO
Comune di Milano

1^a edizione aprile 2019
Tutti i diritti riservati

© Copyright PKE srl

www.pkegroup.it
eMail: info@pke.it

ISBN: 9788894312096

Contents

Preface

As digital technologies, the internet and social media become increasingly integrated into society, a proliferation of digital footprints of human and societal behaviours are generated in our daily lives. All these data provide opportunities to study complex social systems, by the empirical observation of patterns in large-scale data, quantitative modelling and experiments. The social data revolution enables not only new business models but it also provides policy makers with better instruments to support their decisions. This conference aims at stimulating the debate between scholars of different disciplines about the so called "data revolution" in social research. Statisticians, computer scientists and domain experts in social research will discuss the opportunities and challenges of the social data revolution to create a fertile ground for addressing new research problems.

This book includes the abstracts of the papers presented at the second international conference on data science and social research whose authors paid the registration fee.

All abstracts appear in the book as received. Authors are responsible for the entire content and accuracy of their abstracts.

Committees

Scientific committee

Enrica Amaturo, Luigi Fabbris, Carlo Natale Lauro, Paolo Mariani, Enza Messina, Monica Pratesi, Sonia Stefanizzi, Nicola Torelli, Maurizio Vichi.

Programme committee

Biagio Aragona, Francesco Archetti, Carlo Batini, Chiara Binelli, Giovanna Boccuzzo, Federico Cabitza, Furio Camillo, Fabio Crescenzi, Corrado Crocetta, Elisabetta Fersini, Maria Gabriella Grassia, Filomena Maggino, Paolo Mariani, Marina Marino, Andrea Maurino, Enza Messina, Claudio Morana, Matteo Palmonari, Matteo Pelagatti, Sonia Stefanizzi, Rosanna Verde, Giuseppe Vizzari, Biancamaria Zavanella, Emma Zavarrone.

Local organizing committee

Laura Benedan, Antonio Candelieri, Federica Codignola, Alessandra Decataldo, Elisabetta Fersini, Ilaria Giordani, Caterina Liberati, Paolo Mariani, Andrea Marletta, Marcella Mazzoleni, Grazia Murtarelli, Mauro Mussini, Debora Nozza, Matteo Pelagatti, Anisa Rula, Domingo Scisci, Sonia Stefanizzi, Emma Zavarrone, Mariangela Zenga.

Abstracts

INVESTIGATING VACCINE SENTIMENT IN ITALY OVER A PERIOD OF AMBIGUOUS IMMUNIZATON POLICY

Danilo Ajovalasit [1], Veronica Dorgali [2], Angelo Mazza [1], Alberto D'Onofrio [3]
and Piero Manfredi[4]

[1] Dipartimento di Economia e Impresa, Università di Catania,
(e-mail: danilo.ajovalasit@unict.it, a.mazza@unict.it)

[2] Dipartimento di Statistica, Informatica, Applicazioni "G.Parenti", Università di Firenze
(e-mail: veronica.dorgali@gmail.com)

[2] International Prevention Research Institute, (e-mail: alberto.donofrio@i-pri.org)

[2] Dipartimento di Economia e Management, Università di Pisa,
(e-mail:piero.manfredi@unipi.it)

The last 18 months of Italian political life have offered a schizophrenic picture of vaccination programs. Following fell-off of MMR vaccination coverage, a zero-tolerance policy was introduced in Italy, with unvaccinated children not to be allowed to attend school. The new legislation originated a strong debate over the ethical acceptability and the real effectiveness of mandatory vaccination. The new policy was opposed by large strata of the society, including some political parties. The new government that took over after the March 2018 elections, sharply changed the regulation, allowing, under certain conditions, unvaccinated children to school. This research develops a sentiment analysis based on online social media data, in order to investigate whether this confusing phase, with contrasting attitude at the highest level, is creating a disorientation that might eventually lead to distrust in vaccination and to a coverage decline. We will focus on Tweets written in Italian during the past 18 months and containing a set of vaccination-related keywords, to explore which events originated stronger reactions and how they affected vaccination propensity. Furthermore, when available, we will retain the geographic location, to investigate spatial variation within the country.

KEYWORDS: Vaccination choices, sentiment analysis, immunization policy.

References

GETMAN R., HELMI M., & ROBERTS H. 2017. Vaccine Hesitancy and Online Information: The Influence of Digital Networks. *Health Education and Behavior*, https://doi.org/10.1177/1090198117739673.

BELLO-ORGAZ G., HERNANDEZ-CASTRO J., & CAMACHO D. 2017. Detecting discussion communities on vaccination in Twitter, *Future Generation Computer Systems*, http://dx.doi.org/10.1016/j.future.2016.06.032

TRANSITIVITY THRESHOLDS FOR SALO-HAMALAINEN INDEX WHEN THE NUMBER OF ALTERNATIVES IS GREATER THAN THREE

Pietro Amenta [1], Antonio Lucadamo [1] and Gabriella Marcarelli[1]

[1] Department of Law, Economics, Management and Quantitative Methods, University of Sannio, (e-mail: pietro.amenta@unisannio.it, antonio.lucadamo@unisannio.it, gabriella.marcarelli@unisannio.it)

Pairwise comparisons are used for deriving the ranking of preferences in multi-criteria decision problems. Main issue of the pairwise comparisons is the consistency of judgements: they could be not transitive or irrational. It is therefore necessary to measure the level of inconsistency of judgements before deriving a priority vector. Several consistency indices have been proposed in literature to measure the level of inconsistency (e.g. the Salo-Hamalainen index CM_{SH}). They are functions that associate pairwise comparisons with a real number representing the degree of inconsistency in the judgements. Consistency indices and their thresholds may be useful to face cardinal consistency but usually they do not take into account the ordinal consistency (transitivity).

This paper focuses on this issue and proposes a transitivity threshold for the CM_{SH} index providing meaningful information about the reliability of the preferences. If the decision maker is interested in the ordinal ranking of elements and not in the intensity of preferences, then a transitivity threshold represents an important tool for this task: an index value less than the transitivity threshold ensures (with a high probability) that the ranking of preferences is unique while on varying the prioritisation methods, only the intensity of preferences may be different. In this case, even though the index is higher than the consistency threshold, the decision maker may avoid to revise his/her judgments.

KEYWORDS: pairwise comparison matrix, Salo-Hamalainen index, transitivity threshold

References

AMENTA, P. & LUCADAMO, A. & MARCARELLI, G. 2018. Approximate thresholds for Salo-Hamalainen index. *IFAC PapersOnLine*, **51-11**, 1655-1659.

SALO, A. & HAMALAINEN, R. 1997. On the measurement of preference in the analytic hierarchy process. *Journal of Multi-Criteria Decision Analysis*, **6**, 309-319.

INVESTIGATING THE JUDGES' PERFORMANCE IN A NATIONAL COMPETITION OF SPORT DANCE

Laura Anderlucci, Alessandro Lubisco, Stefania Mignani

Department of Statistical Sciences, University of Bologna, Italy
(laura.anderlucci@unibo.it; alessandro.lubisco@unibo.it;stefania.mignani@unibo.it)

Many sports, such as gymnastics, diving, figure skating, etc… use judges' scores to generate a ranking in order to determine the winner of a competition. Judges use some types of rating scale when assessing performances. However, human ratings are subject to various forms of error and bias. Assessment outcomes largely depend upon the set of raters that provide the evaluations.

The aim of this paper is to illustrate how results from Many-Facet Rasch Measurement framework (MFRM) can be used to highlight feedback to the judges about their scoring patterns. The purpose is to analytically detect the bias pattern of the admitted raters. We consider the field of Sport Dance, a discipline receiving an increasing public's interest and passion in recent years. We analyze the data from the national competition of last year in Italy.

KEYWORDS: Rasch model, assessing sport performances, rating variability, sport dance

References

LINACRE, J. M. 2013 Facets computer program for many-facet Rasch measurement, version 3.71.0. Beaverton, Oregon: Winsteps.com

LOONEY, M. A., 2004. Evaluating Judge Performance in Sport, *Journal Of Applied Measurement*, **5(1)**, 31-47.

MYFORD, C. M., WOLFE, E. W., 2003, Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I, *Journal Of Applied Measurement*, ***4*(4),** 386–422.

PRIETO, G., NIETO, E. 2014, Analysis of rater severity on written expression exam using Many Faceted Rasch Measurement, *Psicológica*, ***35***, p.385-397.

COGNITIVE, SOCIAL AND PSYCHOLOGICAL PREDICTORS OF THE GRADUATE'S DISPOSITION TO ENTREPRENEURSHIP

Pasquale Anselmi[1], Daiana Colledani[1], Luigi Fabbris[2] and Egidio Robusti[1]

[1] Department of Philosophy, Sociology, Pedagogy, and Applied Psychology, University of Padua, (e-mail: pasquale.anselmi@unipd.it, daianacolledani@gmail.com, egidio.robusti@unipd.it)

[2] Department of Statistical Sciences, University of Padua (e-mail:luigi.fabbris@unipd.it)

To highlight the factors of psychological capital and locus of control that may influence employability, a survey was carried out on a large sample of graduates from the University of Padua, Italy. The survey was aimed to investigate, among other, the disposition of graduates toward entrepreneurship. Two questionnaires were administered to graduates, one at graduation and another two years later. The questions concerned both the human capital, which refers to the outcomes of graduates' educational investments, other social factors, and a psychometric test to investigate the individual psychological capital (Psycap) and locus of control (LoC). In this work we aim to evaluate if the psychological resources identified by Psycap and LoC scores are related to one's propensity to start own business and to manage in the labour market in such a way to gain a good job in a shorter time than peers. This poses two research questions:
1) which are the psychological factors correlated to entrepreneurship disposition?
2) which are the factors related to the disposition to work abroad?
The psychological dimensions relevant to start own business or to work abroad were identified: the analysis showed that psychological factors can explain the propensity of graduates toward entrepreneurship more accurately than the effects of human and social capital alone.

KEYWORDS: Entrepreneurship, human capital, Social capital, psychological capital,survey on graduates.

SCIENCE MAPPING VIA DYNAMIC TOPIC MODELLING: AN ANALYSIS ON 30 YEARS OF SOCIAL INDICATORS RESEARCH

Massimo Aria[1], Michelangelo Misuraca[2] and Maria Spano[1]

[1] Department of Economics and Statistics, University of Naples Federico II,
(e-mail: massimo.aria@unina.it, maria.spano@unina.it)

[2] Department of Business Administration and Law, University of Calabria,
(e-mail: michelangelo.misuraca@unical.it)

Considerable interest has been devoted in recent years to the quantitative study of the scientific literature, thanks to the availability of online resources (e.g. Web of Science) and the development of effective techniques for performing automatic analyses. The procedures commonly implemented in bibliometric studies are the *performance analysis* and the *science mapping*. The first one aims at evaluating the literature related to a given domain on the basis of bibliographic data. The second one tries to highlight the structural and cognitive patterns of the domain. Mapping techniques frequently refer to textual data analysis. Each domain or theme can be characterised by a set of keywords, assigned by the authors of the publications or by the indexing services.

In this paper, an analysis on the last 30 years of *Social Indicators Research* (SIR) is performed. Founded in 1974, SIR has become one of the leading journals on problems related to the measurement of the different aspects involving the social sphere. Aiming at describing the evolution over time of the SIR themes, here we refer to the *dynamic topic modelling* (DTM) approach (Blei & Lafferty, 2006).

DTM is an extension of the well-known *latent Dirichelet allocation*. It captures the temporal evolution of topics in a sequentially organised collection of documents. The notion of time is included using the meta-data of the documents. By applying DTM to the 1987-2017 SIR collection we show the evolution of keyword-themes distribution, underlying the themes of the collection over the time and tracking how they have changed.

KEYWORDS: bibliometrics, thematic evolution, topic modelling.

References

BLEI, D., & LAFFERTY, J. 2006. Dynamic topic models. *International Conference on Machine Learning*, New York: ACM, 113-120.

ATTRACTIVENESS OF UNIVERSITY DEGREE PROGRAMS: A SOCIAL NETWORK ANALYSIS

Silvia Bacci[1], Bruno Bertaccini[2] and Alessandra Petrucci[2]

[1] Department of Economics, University of Perugia, (e-mail: silvia.bacci@unipg.it)

[2] Department of Statistics, Informatics, Applications "G. Parenti", University of Florence (e-mail: bruno.bertaccini@unifi.it, alessandra.petrucci@unifi.it)

The National (Italian) Register of Students and Graduates (Anagrafe Nazionale Studenti - ANS, in Italian) is an administrative database, established by the Law 170 of 2003 and implemented by the Ministerial Decree 9 of 2004. The aim of this database consists in registering and monitoring the university students enrolled in a degree course. Administrative data (e.g., data about personal characteristics, high school, and students' academic career) is monthly collected from all Italian public and private universities, through an online platform. The ANS database represents a relevant support instrument that is used by local (e.g., universities) and national (e.g., MIUR, ANVUR) decision makers to monitor and evaluate the Italian degree courses. In particular, data from ANS allows the analysis of the national and international mobility of students as well as the choices made by freshmen in terms of degree programs.

The aim of this paper is the analysis of students' migratory movements among universities in order to provide a measurement of the attractiveness of the universities' teaching offer, with a special focus on the masters degree programs. After the bachelor degree, do didactic fields exist where the propensity to move towards a different university is particularly high? Moreover, is this propensity affected by universities' characteristics (e.g., dimension, geographical area)? What are the most attractive universities for each didactic field? The study would answer to these questions applying Social Network Analysis to ANS data related with the enrollments to masters by students that completed the bachelor degree program.

KEYWORDS: anagrafe nazionale studenti, higher education, social network analysis.

References

ENEA, M. 2018. From south to north? mobility of southern Italian students at the transition from the first to the second level university degree. In: Studies in Theoretical and Applied Statistics. Ed. Vichi M. 239-249. Springer. 10.1007/978-3-319-73906-9_22.

KOLACZYK, E. D. 2009. *Statistical Analysis of Network Data. Methods and models*. New York: Springer.

LUKE, D. A. 2015. *A User's Guide to Network Analysis in R*. New York: Springer.

WALKABILITY ASSESSMENT OF URBAN AREAS THROUGH SOCIAL MEDIA DATA MINING

Stefania Bandini[1,2], Andrea Gorrini[1] and Giuseppe Vizzari[1]

[1] Department of Informatics, Systems and Communication, University of Milano-Bicocca
(e-mail: [name.surname]@unimib.it)

[2] Research Center for Advanced Science and Technology, The University of Tokyo

Urban Informatics is an area of application for computer science focused on cities and life in urban areas, with cross-disciplinary contributions from geography, urban planning and social sciences. Urbanization [1], sided by the digital revolution, represents a huge opportunity for this kind of applications aimed at an improvement citizens' quality of life [2]. One of the potential applications of this kind of systems is represented by forms of "bottom-up" evaluations of the status quo, concerning different aspects of the urban texture and supporting decision-making activities. An example of these aspects is walkability [3] (i.e. how comfortable and safe the urban environment is for walking). The computation of indicators describing the characteristics of areas and their usage by pedestrians can be achieved through the exploitation of data from social media, without requiring ad-hoc infrastructures, surveys or observations. This paper will present the results of different analyses on data about the City of Milano (Italy) acquired from different social media and web sources. Acquired metadata were analysed by means of Artificial Intelligence clustering techniques based on the DBSCAN algorithm [4], in order to achieve homogeneous areas characterized by different aspects (i.e. actual activity of inhabitants/tourists, presence of services) rather than by a top down administrative procedure. Results supply useful indications about perceived walkability and they can support the activity of public institutions in the design and planning of the city.

KEYWORDS: urban informatics, walkability, social media data, clustering

References

[1] United Nations (2014). World urbanization prospects: The 2014 revision. UN.

[2] K. Pelechrinis, D. Quercia (2015). In Proceedings of the *24th International Conference on World Wide Web*, pp. 1547-1547.

[3] Speck, Jeff (2013). *Walkable city*. Macmillan.

[4] M. Ester, K. Hans-Peter, S. Jörg, X. Xiaowei (1996). In *Kdd*, vol. 96, pp. 226-231.

EXTRACTION OF CANCER INFORMATION FROM PATHOLOGY CLINICAL RECORDS USING TEXT MINING

Pietro Belloni [14], Giovanna Boccuzzo [1], Stefano Guzzinati [4] Irene Italiano [5],
Bruno Scarpa [1], Carlo R. Rossi [35], Massimo Rugge [24], Manuel Zorzi [4]

[1] Department of Statistical Sciences, University of Padua, Italy
(e-mail: pietro.belloni.1@studenti.unipd.it)

[2] Department of Medicine, University of Padua, Italy

[3] Department of Surgery, Oncology and Gastroenterology, University of Padua, Italy

[4] Veneto Cancer Registry, Padua, Italy

[5] Veneto Oncologic Institute, Padua, Italy

Valuable information is stored in a healthcare record system and over 40% of it is estimated to be unstructured in the form of free clinical text [1]. A collection of pathology records is provided by the Veneto Cancer Registry: these medical records refer to cases of melanoma and contain free text, in particular the diagnosis written by a pathologist and the result of a microscopic and macroscopic analysis of the cancerous tissue. The aim of this research is to extract from the free text the size of the primary tumor, the involvement of lymph nodes, the presence of metastasis, the cancer stage and the morphology of the tumor. This goal is achieved with text mining techniques based on a statistical approach. Since the procedure of information extraction from a free text can be traced back to a statistical classification problem, we apply several data mining models in order to extract the variables mentioned above from the text. A gold-standard for these variables is available: the clinical records have already been assessed case-by-case by an expert. Therefore, it is possible to evaluate the quality of the information that the models are able to extract from the clinical text comparing the result of our procedure with the gold standard.

KEYWORDS: text mining, clinical text mining, data mining, melanoma, pathology records

References

1. H. Dalianis, *Clinical Text Mining: Secondary Use of Electronic Patient Records*. Springer, 2018.

Spatial distribution of Multidimensional Educational Poverty using SAE

Gaia Bertarelli [1], Caterina Giusti [1] and Monica Pratesi [1]

Considering the 2030 Agenda for Sustainable Development, SDG 4 aims to ensure inclusive and equitable quality education and promote lifelong learning opportunities for all". Educational Poverty (EP) is defined as deprivation, for children and adolescents, of the ability to learn, experiment, develop and freely flourish skills, talents and aspirations (Save The Children, 2016). It means being excluded from acquiring the skills needed to live in a world characterized by knowledge-based economy, rapidity and innovation. EP is a latent trait, namely, only indirectly measurable through a collection of observable variables and indicators purposively selected as micro-aspects, contributing to the latent macro-dimension. It is generally measured by the Italian National Statistical Institute by two multidimensional indices, the Educational Poverty Index (EPI) and the Adjusted Mazziotta-Pareto Index (Mazziotta et al. 2016). A problem with these indices is that they are based on direct estimates, which are reliable only at the four broad-areas level, while to intervene on the phenomenon it is important to obtain information at a finer geographical level (NUTS 3 or LAU2). This implies the need of Small Area Estimation (SAE) methods. In this paper we use an adapted version of area-level SAE method proposed by Bertarelli et al. (2018) that uses a Latent Markov Model (LMM) as linking model. In LMMs the characteristic of interest and its evolution in the time is represented by a latent process that follows a discrete Markov chain, usually of first order. Therefore, areas are allowed to change their latent state across time. This model can handle both univariate and multivariate characteristics of interest and can provide a classification of the areas by the intensity of their EP. That is, we can use the area-level data of the single dimension (e.g. literacy skills and competences, PISA OCSE project) or the value of the composite indices.

Keywords: "multidimensional educational poverty, small area estimation, spatial statistics, composite indicators."

References

Bertarelli,G. Ranalli, M.G.; Bartolucci, F; DAl, M. and Solari, F. (2018) Small area estimation for unemployment using latent Markov models Survey Methodology, Vol. 44, No. 2, (in print)

Mazziotta, M., and Pareto, A. (2016). On a generalized non-compensatory composite index for measuring socio-economic phenomena. Social indicators research, 127(3), 983-1003

Save the Children (2016) Liberare I bambini dalla povert educativa. A che punto siamo? available at url: https://www.savethechildren.it

[1] ^{0}Department of Economics & Management, University of Pisa, (e-mail: gaia.bertarelli@ec.unipi.it)

SOCIAL MEDIA DATA FOR SOCIAL INDICATORS: ASSESSING THE QUALITY THROUGH CASE STUDIES

Silvia Biffignandi[1], Annamaria Bianchi[1], and Camilla Salvatore[1-2]

[1] Department of Department of Management, Economics and Quantitative Methods, University of Bergamo (e-mail: silvia.biffignandi@unibg.it, annamaria.bianchi@unibg.it, c.salvatore@studenti.unibg.it)

[2] Department of Economics, Statistics and Marketing, University of Milano-Bicocca (e-mail: c.salvatore4@campus.unimib.it)

KEYWORDS: Big data, social media, Twitter, quality.

As a result of the IT revolution, most of the business or administrative processes and most of the human interactions produce a huge quantity of digital data, better known as "big data" (Laney, 2001). The opportunities are well known: they can be used to answer to new questions, to build new socio-economic indicators, to provide an insight on people's preferences, behaviours, political movements. In particular, the combination of data from multiple sources can provide a better overview of the economic phenomena (Baldacci et al., 2016). Big data is also one of the most discussed topics in Official Statistics and their integration with traditional data sources is a challenging opportunity for the construction of social and economic indicators. It is clear that big data will not replace survey based activity: they can provide complementary, faster and specific information about a topic or they can help to asses unmeasured or partially measured socioeconomic phenomena.

However, there are also new challenges to deal with such as privacy, methodological and especially quality issues. Actually, there is not a precise definition of quality, nor indicators for quality. For this reason we focus on quality issues. Big data substantially differ from survey data, for example, they usually do not correspond to any sampling scheme and are often representative of particular segment of the population, the link between the statistical phenomena of interest and the data is indirect and the inconsistency of data across time and the volatility of the data sources weaken the continuity of the analysis over time. Moreover, in traditional data sources the quality at the origin is checked by the data collector, while, since big data are found data, the quality at the origin is out of the researchers control. Indeed, we should consider big data as an "an imperfect, yet timely, indicator of phenomena in society" (Braaksma and Zeelenberg, 2015).

This paper wants to contribute in defining the quality for big data, building a quality framework for this type of data. Then, we focus on social media data, particularly on Twitter. Social media represents one of the most promising new data sources for social indicators. In the first part of our analysis, we present a review of the possible uses. Then, we discuss the challenges related to the use of this type of data and we assess the overall quality. Quality is a multifaceted concept and, over time, different definitions have been given. According to the Total Survey Quality Framework, the quality is composed by nine dimensions: accuracy, credibility, comparability, usability/interpretability, relevance, accessibility, timeliness/punctuality, completeness and coherence (Biemer, 2010). With regard to the accuracy, the concept of Total Survey Error has been developed and it has been adapted also to Twitter (Hsieh and Murphy, 2017). In our analysis, we try to define a Total Twitter Quality Framework, assessing the different dimensions and we enrich the definition of Total Twitter Error that we analysed in a previous research (Biffignandi, Bianchi and Salvatore, 2018). In particular we discuss the data accessibility issues, whether the definition of quality for survey can be adapted to social media-based analysis and which new dimensions or indicators should be considered. To do this, we analyse different case studies based on the analysis of tweets and users. We then apply text mining, sentiment analysis and topic modelling techniques.

References

BALDACCI, E., BUONO, D., KAPETANIOS, G. KRISCHE, S., MARCELLINO, O., MAZZI, G., & PAPAILIAS, F. (2016). *Big Data and Macroeconomic Nowcasting: from data access to modelling*. Eurostat.

BIEMER, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, **74(5)**, 817-848.

BIFFIGNANDI, S., BIANCHI, A., AND SALVATORE, C. (2018). *Can Big Data provide good quality statistics? A case study on sentiment analysis on Twitter data*. Presented at the "International Total Survey Error Workshop", June 2018, Duke University, North Carolina.

BRAAKSMA, B., ZEELENBERG, K. (2015). ``Re-make/Re-model": Should big data change the modelling paradigm in official statistics?. *Statistical Journal of the IAOS*, ***31*(2)**, 193-202.

HSIEH, Y. P., MURPHY, J. (2017). Total Twitter Error. *Total Survey Error in Practice*, 23-46.

LANEY, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, **6(70)**, 1.

GOOGLE TRENDS AND TWITTER: PREDICTORS OR REACTORS?

AN APPLICATION TO BITCOIN MARKET DETERMINANTS

Biffignandi Silvia[1], Pisanelli Elena[2]

[1] Department of Management, Economics and Quantitative Methods, University of Bergamo.
(e-mail: silvia.biffignandi@unibg.it)
[2] Department of Management, Economics and Quantitative Methods, University of Bergamo.
(e-mail: e.pisanelli@studenti.unibg.it)
[2] Department of Economic and Social Sciences, Collegio Carlo Alberto.
(e-mail: elena.pisanelli@carloalberto.org)

Starting from the assumptions that information has deep effects on individuals' socio-economic behavior, and that social networks' data provide huge amounts of information about individuals' socio-economic choices, this work applies social networks' data analysis, using Google Trends and Twitter's data, on a new topic yet statistically unexplored: Bitcoin. We tried to understand whether Google searches and Tweets act as predictors or reactors of market price across Bitcoin exchange and Bitcoin trade volumes. Using global data coming from Google Trends' queries and Twitter, we tried to give a clear definition of Google searches and Twitter's role, making use of Pearson correlation coefficient and Granger-causality test. Our results provide evidence that searches in Google act as predictors of Bitcoin market price, while Twitter acts as reactor, responding to the change in market price after its occurrence. A key role in this mechanism is played by Bitcoin's popularity. An attention allocation process, driven by Google searches, is originated by new investors, seeking for potential investment opportunities in Bitcoin. The process affects market determinants and is boosted by cumulative share of information through Google and Twitter. To check for our findings' robustness, we built up words clouds to confirm the strong correlation among Google Trends, Twitter and Bitcoin market determinants. We collocate this work as part of the studies on socio-economic behavior of individuals.

KEYWORDS: Big Data, Bitcoin, Google Trends, Sentiment Analysis, Twitter.

References

J. BOLLEN, H. MAO, AND X. ZENG. 2011.Twitter mood predicts the stock market. *Journal of Computational Science.,* **2.1,** 1–8.

J. MONDRIA, T. WU, AND Y. ZHANG. 2010. The determinants of international investment and attention allocation: Using internet search query data. *Journal of International Economics.,* **82.1**, 85–95.

T. RAO AND S. SRIVASTAVA. 2012. Analyzing stock market movements using twitter sentiment analysis. *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining* (ASONAM 2012). IEEE Computer Society. 119–123.

THE ROLE OF PARTIES AND MEDIA FOR INCOME INEQUALITY PERCEPTIONS. EVIDENCE FROM THE UNITED STATES AND THE UNITED KINGDOM

Chiara Binelli[1] and Paul Matthew Loveless[2]

[1] University of Milano-Bicocca (e-mail: chiara.binelli@unimib.it)

[2] Center for Research and Social Progress, (e-mail: m2loveless@gmail.com)

We exploit the detailed individual-level data on media consumption and party affiliation of both the British Election Study internet panel and the American National Election Study to estimate a comprehensive model of income inequality perceptions. By modeling media and party choices as both independent covariates and functions of individuals' ideological self-location, we find that party choices and normative orientations are central to inequality perceptions. Further, while party affiliations in both countries are determined by individual ideology, the impact of parties on perceptions comes primarily from conservative parties. For media, in addition to scattered independent effects in the US, there is consistent cross-national evidence for 'selective exposure' effects in which individual political ideology correlates with media choice, which in turn impacts on inequality perceptions. Thus, the inclusion of parties and media unveils a potentially broader role of political ideology as a crucial determinant of inequality perceptions than previously identified.

KEYWORDS: Perceptions of Income Inequality, Media Effects, Public Opinion, United States, United Kingdom

RISKINESS OF ITALIAN FIRMS IN THE POST-CRISIS PERIOD: AN OUTLOOK THROUGH FINANCIAL RATIOS

Matilde Bini[1], Lucio Masserini[2] and Alessandro Zeli [3]

[1] Department of Statistics, European University of Rome, (e-mail: Matilde.Bini@unier.it)

[2] Department of Economics and Management, University of Pisa,
(e-mail: lucio.masserini@unipi.it)

[3] Division for data analysis and economic, social and environmental research, ISTAT,
(e-mail: alessandro.zeli@istat.it)

The paper aims to analyze riskiness of Italian firms in productive sectors (excluding financial and public sector) by means of a set of financial ratios calculated on the basis of annual financial reports collected by Chamber of Commerce for Italian limited firms. This large administrative source (integrated with business demographic information coming from ISTAT business register ASIA) enables to elaborate data concerning the limited enterprises in Italy for a period lasting from 2008 to 2014 (the period after the Big Recession). The limited firms represent a large and significant part of the economy. Several areas are identified as relevant in building riskiness indicator: leverage ratios, efficiency measures, performance measures, and liquidity measures.
The riskiness level is analyzed by means of latent variables models, therefore a "synthetic'' indicator summarizing the data information and representing the riskiness of Italian firm in the period is built.

KEYWORDS: bankruptcy risk, firm riskiness, latent variables models, financial ratio analysis.

References

ALTMAN, E., MARCO, G., & VARETTO, G. 1994. Corporate distress diagnosis: comparison using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking & Finance*., **18**, 505–529.

BOTTAZZI, G., GRAZZI, M., SECCHI, A., & TAMAGNI, F. 2011. Financial and economic determinants of firm default. *Journal of Evolutionary Economics*., **21**, 373–406.

DAVIES, S., RONDI, L., & SEMBENELLI, A. 2001. S.E.M and the changing structure of EU manufacturing 1987–1993. *Industrial and Corporate Change*., **10**(1), 37–75.

ZELI, A., & MARIANI, P. 2009. Productivity and profitability analysis of large Italian companies: 1998–2002. *International Review of Economics*, **56**, 175–188.

ZELI, A. 2014. The Financial Distress Indicators Trend in Italy. An Analysis of Medium–size Enterprises. *Eurasian Economic Review.*, **4**(2), 199-221.

MEDIA BIAS AND CRIME PERCEPTION

Riccardo Borgoni[1], Daniele Gualtieri[1] and Alessandra Michelangeli[1]

[1] Department of Economics, Statistics and Marketing, University of Milano-Bicocca, (e-mail: Riccardo.borgoni@unimib.it, alessandra.michelangeli@unimib.it)

It is commonly accepted that violent crimes are unevenly distributed in time and space. Spatial clusters of criminal events also tend to reappear regularly or remain stable over time. This suggests that crimes spread through local environments via a contagion-like process (Johnson, 2008) due to a number of reasons. For instance, a spatial concentration of socio-economic disadvantages, a local low enforcement control or the presence of environmental facilitators can be potential causes of such clustering. This paper is primarily intended as a contribution in terms of empirical knowledge of the systematic behaviour of spatio-temporal crime patterns within a modern European city. We consider 1,678 violent crimes committed in the city of Milan between 2010-2012, and reported in national and local newpapers. Although crime news presents a partial and possibly biased image of the crimes actually committed, they are an important indicator for both spatial perception and spatial fear of crime. Hence, this piece of research manages to provide a map of the crime risk perceived by the city's inhabitants on the basis of the violent crimes reported in the newspapers. On the modelling ground, it is assumed that crime occurrences conform to a self exciting non homogeneous spatio-temporal point process (SENHPP). SENHPP have been proposed to model earthquake occurrences since earthquakes are well known to increase the risk of subsequent earthquakes, or aftershocks, near the location of the initial event. SENHPP have also been suggested to model burglaries (Mohler et al.2011) using US data. However, to the best of our knowledge, this paper is the first attempt to apply this approach to European media data. More specifically, we assume that the conditional intensity of the SENHPP can be modelled via the combination of three functions representing a spatial background component, a temporal background component and a spatio-temporal interaction component that takes into account the interaction of each crime with previous (triggering) events. The model is estimated by coupling a Monte Carlo declustering algorithm that accounts for the self exciting nature of crimes with variable bandwidth kernel estimation. We show how this approach can be usefully used to predict crime occurrences in a given time span to provide a picture of the perceived risk of crime and also how it can be adopted to support authorities in monitoring the urban territory.

KEYWORDS: crime mapping, crime perception, self-exciting spatiotemporal point process, mass media

References

JOHNSON, S. D. (2008). REPEAT BURGLARY VICTIMISATION: A TALE OF TWO THEORIES. JOURNAL OF EXPERIMENTAL CRIMINOLOGY 4 (3), 215-240.

MOHLER, G., M. SHORT, P. BRANTIGHAM, F. SCHOENBERG, AND G. TITA (2011). SELF-EXCITING POINT PROCESS MODELING OF CRIME. JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION 106 (493), 100-108.

The Evolution of Inequality of Opportunity in Germany: A Machine Learning Approach[*]

Paolo Brunori[†] Guido J. Neidhöfer[‡]

October 29, 2018

Abstract

Following Roemer (1998) we assume that valuable outcomes individuals obtain are the joint result of their circumstances and the effort they exert. The two components can be separated following a two-step procedure. First, identifying types, i.e. groups of individuals characterized by the same circumstances beyond individual control. Second, measuring the degree of effort exerted by each individual. Inequality of opportunity is then inequality between individuals exerting the same degree of effort but belonging to different types. Roemer's approach is not very frequently adopted by empirical economists because it requires to explicitly model the role of effort. The measurement of effort is a challenging empirical exercize and it has often diverted the focus of economists toward simpler measure of inequality of opportunity. In what follows we show how measures of inequality of opportunity fully consistent with Roemer's approach can be straightforwardly estimated adopting a machine learning approach. This approach uses both regression trees, to identify Romerian types, and polynomial approximation, to estimate the degree of effort exerted by individuals. Our method has two important advantages: first, it allows to relax a number of arbitrary assumptions otherwise necessary to measure effort. Second, opportunity trees can be displayed graphically and are easily interpreted, allowing an intuitive representation of the evolution of inequality of opportunity. We illustrate our approach measuring inequality of opportunity in Germany from 1990 to 2016 taking advantage of information contained in 26 waves of the German Socio-Economic Panel.

[*] We are grateful to Torsten Hothorn for useful suggestions. Any errors remain our own.

[†] University of Florence, Dipartimento di Scienze per l'Economia e l'Impresa, Via delle Pandette 32 - 50127 Firenze, Italy, paolo.brunori@unifi.it.

[‡] ZEW Centre for European Economic Research, guido.neidhoefer@zew.de.

MAPPING BRAND PUBLICS' SOCIAL IMAGINARIES ON INSTAGRAM: HOW TO USE BIG DATA FOR EXPLORING CONSUMER CULTURE

Alessandro Caliandro[1] and Guido Anselmi[2]

[1] School of Management, University of Bath
(e-mail: a.caliandro@bath.ac.uk)

[2] Department of Sociology and Social Research, University of Milano-Bicocca
(e-mail: guido.anselmi@unimib.it)

In the last decade consumer research has focused on online communities as privileged digital sites for consumers and admirers of brands to develop social bonds and shared identities. Recently researchers have begun to debate the relevance of the concept of 'community' for understanding brand-related communication on social media such as Facebook, Twitter, and Instagram. In a recent article Arvidsson and Caliandro (2016) proposed the notion of brand public as a label to capture current modes of online participation, characterised by a lack of interaction and practices of co-creation of common social imaginaries around brands. While Arvidsson and Caliandro have elaborated a new analytical category endowed with an appreciable heuristic force, we feel that they have yet to propose a convincing methodology for framing the processes through which disconnected social media users are actually able to co-create common imaginaries around brands.

To this purpose we set a research project aimed at investigating the existence of such social imaginaries as well as their semantic nature. The research project is based on a dataset of 1,100,000 Instagram photos marked with the hashtag of 6 prominent global brands (#Starbucks, #LouisVuitton, #Zara, #McDonalds, #Smirnoff, #GreyGoose), which we analysed by taking advantage of software of image recognition (Google Vision API). Combining automated and manual analysis we came up with the following results: a) a set of procedures and techniques for segmenting consumers belonging to a brand public; b) the presence of a (peculiar) repetitive visual pattern in the photos posted by consumers belonging to brand publics. This study amounts to be relevant to better understand the new forms of fluid identity and values (Bardhi and Eckhardt 2017) consumers develop around brands on social media.

KEYWORDS: ‘brand public’, ‘Instagram’, ‘digital method’, ‘consumer culture’, ‘automated visual analysis’.

References

ARVIDSSON, A. & CALIANDRO, A. 2016. Brand public. *Journal of Consumer Research*, **42.5**, 727-748.

BARDHI, F. & ECKHARDT, G.M. 2017. Liquid consumption. *Journal of Consumer Research*, **44.3**, 582-597.

ANALYSIS OF TWO-WAY ORDINAL CONTINGENCY TABLES FOR SOCIAL RESEARCH

Ida Camminatiello [1], Antonello D'Ambra[1] and Luigi D'Ambra[2]

[1] Department of Economics, University of Campania, L. Vanvitelli (e-mail: ida.camminatiello@unicampania.it, antonello.dambra@unicampania.it)

[2] Department of Economics, Management and Institutions, University of Naples, Federico II (e-mail: dambra@unina.it)

In the social research the variables often consist of ordered categories. When the row and column variables of a contingency table both are on ordinal scale, several techniques (Beh, 1997) have been proposed, most of which are based on the partition of Pearson's chi-squared statistic. Recently doubly ordered cumulative correspondence analysis (D'ambra, Beh, Camminatiello, 2014) has been proposed by partitioning Hirotsu's chi-squared statistic (Hirotsu, 1994). The association in these tables can be also described by using various types of odds ratios, among which, global odds ratios (Agresti, Coull, 2002).

In this contribution we propose a modification of the above doubly ordered cumulative correspondence analysis based on the logarithms of the elements of the doubly cumulative table obtained by collapsing row and column classifications into dichotomies. By means of this generalization we can compute and represent the global odds ratios in the two-dimensional plot.

KEYWORDS: ordered categories, global odds ratios, doubly cumulative table.

References

AGRESTI, A., & COULL, B.A. 2002. The analysis of contingency tables under inequality constraints. *J. Statist. Plann. Inference*, 107, 45–73.

BEH, E. J. 1997. Simple correspondence analysis of ordinal cross-classifications using orthogonal polynomials. Biometr. J. 39:589–613.

D'AMBRA, L., BEH, E. J., & CAMMINATIELLO, I. 2014. Cumulative Correspondence Analysis of Two-Way Ordinal Contingency Tables. *Communications in Statistics - Theory and Methods*, 43 (6), 1099-1113.

HIROTSU, C. 1994. Modelling and analysing the generalized interaction. *Proc. Third IEEE Conf. Control Applic.*, 2, 1283–1288.

GLOBAL OPTIMIZATION OF A MACHINE LEARNING BASED FORECASTING PIPELINE

Antonio Candelieri[1], Riccardo Perego[1] and Francesco Archetti[1,2]

[1] Department of Computer Science, Systems and Communication, University of Milano-Bicocca,
(e-mail: antonio.candelieri@unimib.it, Riccardo.perego@unimib.it)

[2] Consorzio Milano Ricerche,
(e-mail: archetti@milanoricerche.it)

Global Optimization, especially Bayesian Optimization, has become the tool of choice in hyperparameter tuning and automatic algorithm configuration in the Machine Learning community. This paper presents a Bayesian Optimization framework for the optimal design of a forecasting pipeline based on time series clustering and Artificial Neural Networks. The software environment R has been used with the mlrMBO package. Random Forest has been adopted as probabilistic surrogate model, due to the nature of decision variables (i.e. conditional and discrete hyperparameters) in the pipeline design. Both Expected improvement and Lower Confidence Bound were used as acquisition function and results are compared.

The computational results, on a benchmark and a real-world dataset, show that even in a complex search space, with integer, categorical and conditional variables, the proposed Bayesian Optimization framework is an effective solution, with Lower Confidence Bound requiring a lower number of function evaluations than Expected Improvement to find the same optimal solution.

KEYWORDS: Bayesian optimization, machine learning, hyperparameters optimization, automatic algorithm configuration.

References

MALA-JETMAROVA, H., SULTANOVA, N., & SAVIC D. 2017. Lost in Optimization of Water Distribution Systems? A literature review of system operations, *Environmental Modelling and Software*, **93**, 209-254.

SNOEK, J., LAROCHELLE, H., & ADAMS, R.P. 2012. Practical Bayesian Optimization of Machine Learning Algorithms, *arXiv:1206.2944 [stat.ML]*.

THORTON, C., HUTTER, F., HOOS, H.H., & LEYTON-BROWN, K. 2013. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. *In: Proceedings of ACM SIGKDD. 847–855*.

FEURER, M., KLEIN, A., EGGENSPERGER, K., SPRINGENBERG, J., BLUM, M., HUTTER, F. 2015. Efficient and robust automated machine learning. *In Advances in Neural Information Processing Systems, 2962-2970*.

APPROXIMATE DYNAMIC PROGRAMMING FOR PUMPS SCHEDULING OPTIMIZATION IN URBAN WATER DISTRIBUTION SYSTEMS

Antonio Candelieri[1], Riccardo Perego[1] and Ilaria Giordani[1]

[1] Department of Computer Science, Systems and Communication, University of Milano-Bicocca, (e-mail: antonio.candelieri@unimib.it, riccardo.perego@unimib.it, ilaria.giordani@unimib.it)

Urban water distribution networks are becoming complex Cyber-Physical Systems generating huge amounts of data from smart metering and flow/pressure monitoring. Thus, strategies for the operations optimization must adapt to the new data rich environment where decision making must be near real time. In this paper, an Approximate Dynamic Programming (ADP) approach is proposed to deal with the Pump Scheduling Optimization (PSO) problem, aimed at identifying an optimal schedule for the pumps to minimize associated energy costs while satisfying operational constraints (e.g. water demand, pressures within a given range, and reservoir levels within some pre-specified min-max range). More precisely. Q-Learning, one of the ADP algorithms, well known in the Reinforcement Learning community, is used. Traditional optimization strategies fail to capture the value hidden in real time data, and usually require knowing the water demand in advance or, at least, to have a reliable and accurate forecast. On the contrary, ADP learns a strategy to decide which pumps must be activated, at each time step, depending on the online observation of the system. Results on the Anytown benchmark proved that ADP is robust with respect to uncertainty on the water demand and able to deal with real time data, with no distributional assumptions or demand forecasts needed.

KEYWORDS: approximate dynamic programming, reinforcement learning, pump scheduling optimization, water distribution networks.

References

MALA-JETMAROVA, H., SULTANOVA, N., & SAVIC D. 2017. Lost in Optimization of Water Distribution Systems? A literature review of system operations, *Environmental Modelling and Software*, **93**, 209-254.

POWELL, W.B. 2007. Approximate Dynamic Programming: Solving the Curses of Dimensionality. *John Wiley and Sons*.

FRACASSO, P.T., BARNES, F.S., COSTA, A.H.R. 2013. Energy cost optimization in water distribution systems using Markov Decision Processes. *International Green Computing Conference Proceedings*, Arlington, 1-6.

MODELLING HOUSING MARKET CYCLES IN GLOBAL CITIES

Alessandra Canepa[1] Emilio Zanetti Chini[2] Huthaifa Alqaralleh[3]

[1] Department of Economic and Statistics Cognetti De Martiis, University of Turin, and Department of Economics and Finance, Brunel University London. (e-mail: Alessandra.Canepa@unito.it)

[2] Department of Economics and Management, University of Pavia. (e-mail: emilio.zanettichini@unipv.it)

[3] Department of Economics, Business and Finance, Mutah University. (e-mail: huthaifa89@mutah.edu.jo)

In this paper we consider the dynamic features of house prices in metropolises that are characterized by high degree of internationalization. Using a generalized smooth transition model we show that the dynamic symmetry in house price cycles is strongly rejected for the housing markets taken into consideration.

KEYWORDS: house price cycles, dynamic asymmetries, nonlinear models.

JEL CLASSIFICATION: C10, C31, C33.

DIVORCE IN ITALY: A TEXTUAL ANALYSIS OF CASSATION JUDGMENTS

Rosanna Cataldo [1], Maria Gabriella Grassia [2], Marino Marina[2], Rocco Mazza[2], Vincenzo Pastena[3]

[1] University of Naples ”Federico II”, Department of Economics and Statistical Science, (e-mail: rosanna.cataldo2@unina.it)

[2] University of Naples ”Federico II”, Department of Social Sciences, (e-mail: mgrassia@unina.it, mari@unina.it, rccmazza@gmail.com)

[3] Studio legale Pastena, (e-mail: avv.vincenzo.pastena@gmail.com)

The paper aims to study the social developments in the phenomenon of divorce in Italy within the judgments of cassation, a reference point for national jurisprudence, and to integrate these results both with current demographic trends and with an analysis conducted on web community that collect experiences of interruption of marriages. The study will carried out using text mining tools, for the analysis of judgments and topics collected in web communities, with particular attention to tool for automatic collection of texts from web and the data pre-treatment phase. Analysis and integration of data of different type will allow a better understanding of the object of study. In fact, the institutional and juridical dimension is first inserted into a demographic framework that allows a contextualization and then the study of the web community will offer a comparison with a relational plan, based on the testimonies of those who live the divorce.

KEYWORDS: Divorce, text mining, demography, web communities.

References

BOLASCO S. 2005. Statistica testuale e text mining: alcuni paradigmi applicativi, *Quaderni di statistica*, vol.**7**.

HOFMANN M., & CHISHOLM A. 2016. *Text Mining and Visualization: Case Studies Using OpenSource Tools*, Chapman & Hall/CRC.

ISTAT, 2016. *Le trasformazioni demografiche e sociali: una lettura per generazione*

MARET, P., & RAJENDRA A., & LAURENT V., 2016. *Web Communities in Big Data Era*. Proceedings of the 25th International Conference Companion on World Wide Web, p. 945-947, Monteral, Canada.

DIFFERENT SOURCES OF DATA FOR THE SUSTAINABILITY

Rosanna Cataldo[1], Maria Gabriella Grassia[2], Marina Marino[2] and ViktoriiaVoitsekhovska[3]

[1] University of Naples Federico II Department of Economics and Statistical Science,
(e-mail: rosanna.cataldo2@unina.it)

[2] University of Naples Federico II Department of Social Science,

(e-mail: mgrassia@unina.it, mari@unina.it)

[3] Department of economics of enterprise, Lviv Polytechnic National University, Institute of economics and management, (e-mail: viktoriia.v.voitsekhovska@lpnu.ua)

Social media have become an emerging phenomenon due to the huge and rapid advances in information technology. People are using social media on daily basis to communicate their opinions with each other about wide variety of subjects and general events. Social media communications include Facebook, Twitter, and many others.

Through Social media, in this work, we examine people's feelings to a phenomenon, sustainability, that is the greatest challenge of our generation.

Specifically, we collected comments on principal social networks used from people and the textual information has been analyzed through techniques of Text Mining and Network Analysis in order to detect some important structures of people communication, understanding their mood about this concept.

KEYWORDS: Sustainability, Text Mining, Social Network Analysis

MAPPING TRENDING TOPICS IN SOCIAL RESEARCH METHODS

Maria Carmela Catone[1], Paolo Diana[1], Giuseppe Giordano [1], and Pierluigi Vitale[1]

The Social Research Methods is a scientific domain facing a transformation mainly due to: *i*) new challenges induced by the complexity of the contemporary society (Castellani, Hafferty, 2009), *ii*) the availability of different kind of data and new data sources (e.g. data derived by social media, micro blogging platforms, tagging practices) and, *iii*) data analytics tools that merge Statistics and Information Science techniques. As a result, scholars need to develop new approaches to deal with the design and all the consequent phases of the research. Our aim is to analyze the current literature looking for the trending topics and trying to delineate the leading edges of the Social Research Methods. In this paper we explore the scientific literature related to this domain, building a bibliography database collected in the period 2011-2017 and published in the first five Journals having the maximum Impact Factor in the field. The database is characterized by different relevant bibliographic attributes, such as: Authors, Years, Keywords and the Abstracts as textual data. The main purposes is to explore such data collection and analyze the evolution of traditional themes as well as the presence of new topics. Firstly, we investigate the keywords in order to build relational patterns linking documents and scholars in the scope of network analysis (Wasserman, Faust, 1994). At this end, the bipartite network of documents and keywords will be defined and investigated. Afterwards, textual data arranged in a lessical table provided by documents and Abstracts' lemmas, will be analyzed by means of textual data mining (Lebart et Al., 1998). The network and the textual data will be then joint analyzed to focus on the visualization of thematic patterns and the existence of *structural holes* leading to innovation and potential new scenarios for Social Research Methods.

KEYWORDS: bibliographic data, co-author network, social network analysis, textual data mining

References

CASTELLANI, B., & HAFFERTY, F. W. 2009. *Sociology and complexity science: a new field of inquiry*. New York: Springer.

LEBART L., SALEM A., BERRY L. 1998. Correspondence Analysis of Lexical Tables. *Exploring Textual Data. Text, Speech and Language Technology*, **4**. Springer, Dordrecht.

WASSERMAN, S., & FAUST, K. 1994. *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge: Cambridge University Press.

[1] [0] Department of Social, Political and Communication Sciences, University of Salerno, e-mail: mcatone@unisa.it, diana@unisa.it, ggiordano@unisa.it, pvitale@unisa.it

COMPOSITE INDICATORS OF THE SOCCER PLAYERS' PERFORMANCE INDICES

Enrico Ciavolino[1], Maurizio Carpita[2] Paola Pasca[1]

[1] Department of History, Society and Human Studies (University of Salento, Lecce, Italy), (e-mail: enrico.ciavolino@unisalento.it, paola.pasca@unisalento.it)

[2] Department of Economics and Management (University of Brescia, Italy), (e-mail: maurizio.carpita@unibs.it)

Most of the prolific data science and statistical competitions occur in the context of soccer matches. Performance variables are used as predictors or combined with other relevant information to build new performance indicators that may help to predict the result of a match (win, draw and loss of the home team). In the field of soccer, soFIFA experts' of EA Sports are considered the leading authority for what concerns the evaluation of players' performance: they state the overall sport performance consists of some dimensions, each of which in turns incorporates more specific skills to be developed and mastered by players on the soccer field. However, the statistical properties of the SoFIFA experts' indicators have never been explored and assessed from the statistical point of view. This work aims at pursuing this goal, through both a non-supervised and a supervised modelling approach.

KEYWORDS: soccer indicator modelling, performance indicator assessment, supervised and non-supervised approach

MODELLING AND TESTING ON MULTIVARIATE LONGITUDINAL DATA FOR NESTED DESIGN WITH APPLICATION TO PLAYER-BY-PLAYER BASKETBALL ANALYTICS

Livio Corain[1], Luigi Salmaso[1]

[1] Department of Management and Engineering, University of Padova, Italy,
(e-mail: livio.corain@unipd.it, luigi.salmaso@unipd.it)

Suppose that on several subjects/individuals, each one belonging to a mutually exclusive group of interest, we may observe multiple times a multivariate repeated measure in which each univariate component can be either binary or numeric or ordered categorical. By modelling this kind of nested design as a longitudinal linear fixed effect model and by using the union-intersection approach with the emphasis placed on the ranking of location effects, the goal of the present paper is proposing a multivariate testing approach for doing inference on both between and within groups analysis. Our approach may be effective for handling with some real problems in research fields such as behavioural and social sciences as well as in sport analytics. Via a Monte-Carlo simulation study we investigated the properties of the proposed testing and ranking methodology and we proved its validity under different random distributions. Finally, by using play-by-play basketball data, we present an application to player-based data sport analytics.

KEYWORDS: permutation tests, *p*-value combination, union-intersection principle.

References

ARBORETTI, R., BONNINI, S., CORAIN, L., SALMASO, L. 2014. A Permutation Approach for Ranking of Multivariate Populations, *Journal of Multivariate Analysis*, DOI: 10.1016/j.jmva.2014.07.009, **132**, 39–57.

BONNINI, S., CORAIN, L., MAROZZI, M., SALMASO, L. 2014. *Nonparametric Hypothesis Testing: Rank and Permutation Methods with Applications in R*, Chichester : Wiley.

CORAIN, L., ARBORETTI, R., CECCATO, R., RONCHI, F., SALMASO, L. 2018. Testing and Ranking on Round-Robin Design for Data Sport Analytics with Application to Basketball, *Statistical Modelling*, to appear.

DETECTION OF MUTIMEDIA SEXIST CONTENTS

Silvia Corchs, Elisabetta Fersini and Francesca Gasparini
Department of Informatics Systems and Communication, University of Milano-Bicocca, (e-mail: silvia.corchs@unimib.it,
elisabetta.fersini1@unimib.it,
francesca.gasparini@unimib.it)

Multimedia data, especially shared in online social media, are often based on visual and/or textual contents frequently showing women as subjects. It is highly probable, especially in advertisement, or meme that a woman is portrayed in a highly sexualized manner (Zimmerman and Dahlberg 2008). Both the image and the accompanying text can encode several forms of sexism. Among them, we can highlight the most prevalent ones (Poland 2016): i) Stereotype: women are typically portrayed as a good wives mainly concerned with tasks of housekeeping; ii) Objectification: women are presented as sex objects; iii) Dominance: women are depicted as physically or mentally dominated by men. To address the problem of automatic detection of sexist multimedia contents, we have first created different labeled databases of texts and related images. In particular we have created a manually labeled database of sexist and non sexist advertisements, composed of two main datasets: a first one containing 423 advertisements with images that have been considered sexist (or non sexist) with respect to their visual content, and a second dataset comprising 192 advertisements labeled as sexist and non sexist according to visual and/or textual cues. We have also created a database of about 800 memes that have been evaluated using a crowdsourcing web platform and labeled according to visual and/or textual cues. Moreover, taking into account the promising classification results of multimodal approaches, we investigate several approaches where both visual and texture features are analyzed to detect multimedia sexist contents.

KEYWORDS: sexism, image classification, multimodal classification, text analysis

References

ZIMMERMAN, A., & DAHLBERG,J. 2008. The sexual objectication of women in advertising: A contemporary cultural perspective. *Journal of Advertising Research*, **48(1)**, 7179.

POLAND, B. 2016. *Haters: Harassment, Abuse, and Violence Online*. U of Nebraska Press.

BAYESIAN NETWORKS TO DISCOVER SIMILARITIES AMONG SUBSETS. COMPARING EUROPEAN COUNTRIES

Rosario D'Agata[1], Simona Gozzo[1] and Anna Maglia[1]

[1] Department of Social and Political Sciences, University of Catania,
(rodagata@unict.it, simonagozzo@yahoo.it, angela.maglia@outlook.it)

Bayesian Networks, whose knowledge and usage is mainly spread in the medical field, allow reducing complexity (Brogini and Slanzi, 2010) even when data collection comes from surveys. In these contexts, where many variables representing different dimensions are at stake, creating paths that highlight their connections may be useful for showing similarities among subsets. Many packages allow to create Bayesian Networks, but the one that is considered the best in terms of efficiency seems to be bnlearn (Scutari, 2010). Bnlearn is a R package that, through learning algorithms and conditional independence tests, provides network scores that enable to build the best network solutions. In this work, based on the European Social Survey (ESS) data, we aim at comparing different countries in relation to their welfare models looking at how their variables move, especially in terms of trust and participation. Furthermore, we will attempt to find out paths similarities between countries in terms of cohesive mechanisms (Green *et al.*, 2009).

KEYWORDS: Bayesian network, bnlearn, ESS, welfare

References

BROGINI, A., & SLANZI, D., 2010. On using Bayesian networks for complexity reduction in decision trees. *Statistical Methods and Applications*, **19**, 127-139.

GREEN, A., JANMAAT, J. G., HAN, C., 2009. Regimes of Social Cohesion. *Published by the Centre for Learning and Life Chances in Knowledge Economies and Societies at: http://www.llakes.org.uk.*

SCUTARI, M., 2010. Learning Bayesian Network with bnlearn R package. *Jornal of Statistical Software,* **35**, issue 3.

QUALITY ASPECTS WHEN USING MOBILE PHONE DATA IN OFFICIAL STATISTICS

Fabrizio De Fausti[1], Roberta Radini[1], Luca Valentino[1] and Tiziana Tuoto[1]

[1] ISTAT, Italian National Statistical Institute,
(e-mail: defausti@istat.it,radini@istat.it,luvalent@istat.it, Tuoto@istat.it)

KEYWORDS: mobile phone data, population size estimation, human mobility

The mobile phone data have recently shown great potentialities in official statistical production, even if they are generated for very different purposes. The statistical uses of these data need to take into account the generation model of the data. Actually different kinds of data are generated by the extremely complex interaction between each mobile device and the network, some of them being only temporarily stored. These data enter into a cascade of larger systems for several internal purposes, e.g. to bill. Obviously, different data show different potentialities, however, a common point is the need to pre-process the data in order to generate so-called statistical microdata, suitable for further statistics procedures.

The main advantages of using mobile phone data are related to their granularity, the mobile operator network is widespread on the land, and the timeliness compared to other traditional sources, potentially the data are available in real-time. However, it is necessary to find a compromise between the processing cost, the quality requirements of the statistical output, and the privacy constraints. A crucial role is played by the international phone standards that can be the basis for definitions and statistical metadata of mobile data. This would allow consistent statistics, comparable over the time and between regions and countries. The definition of efficient metadata allows combining and making integrated use of data from different sources.

This paper deals with the usages of mobile phone data in official statistics, with applications to population estimates and human mobility statistics. We highlight the advantages of the use of mobile phone data as well as some quality aspects that need further investigation to be aligned with standard official statistics production.

References

DEVILLE P, LINARD C, MARTIN S, GILBERT M, STEVENS FR, GAUGHAN AE, BLONDEL D AND TATEM AJ. 2014. Dynamic population mapping using mobile phone data, *PNAS*, 111, 45

FURLETTI B, TRASARTI R, CINTIA P AND GABRIELLI L. 2017. Discovering and Understanding City Events with Big Data: The Case of Rome, *Information*, 8,74.

FUNDAMENTAL PRINCIPLES OF OFFICIAL STATISTICS *http://www.unece.org/stats/archive/docs.fp.e.html*

A TOURIST TOUR PLANNING FOR A SMART TOURISM SYSTEM IN CALABRIA

Annarita De Maio[1], Francesco Santoro[2,3] and Antonio Violi[2,3]

[1] Department of Information Systems and Telecommunications, University of Milano Bicocca,
(e-mail: annarita.demaio@unimib.it)

[2] ITACA s.r.l., Rende (CS), Italy
(e-mail: santoro@itacatech.it, violi@itacatech.it)

[3] Department of Mechanical, Energy and Management Engineering, University of Calabria

We introduce the description and development phase of a real Smart Tourism System born in the context of a national project for building a decisional support system that helps the tourists in the planning phase of their trip. The SmartCal (Smart Tourism in Calabria) project aims at delivering a strategy for tourism development in a region, based on the preferences and the needs of modern tourists. The system is developed considering different aspects: evolvement of the decision makers and the stakeholders of the tourism sector, analysis of Point of Interest (POI) with their relationship with the transport systems and infrastructures, designing of a proactive tourist tour planner and static and dynamic profiling of the users. The output of the project will be a platform and a smartphone application that offers many functionalities to the user. The project core is the development of the proactive tourist tour planner, that builds an itinerary for visiting a set of points of interest, considering the user preferences learned from the social networks analysis.

The tourist tour planning engine is based on ad hoc method for a specific version of the Orienteering Problem (OP). This is a routing problem which has lots of applications in logistics, tourism and defence. Given a set of nodes, POIs, the decision support model aims at designing a tour visiting a subset of POIs. The objective of the problem is to maximize the total score while the total travel time and the total cost of the route do not exceed some predefined thresholds. In this work, we present a genetic algorithm framework combined with some local search operators to deal with the analysed problem.

KEYWORDS: smart tourism, orienteering problem, genetic algorithm.

WEB-BASED DATA COLLECTION AND QUALITY ISSUES IN CO-AUTHORSHIP NETWORK ANALYSIS

Domenico De Stefano[1], Vittorio Fuccella[2], Maria Prosperina Vitale[3] and Susanna Zaccarin[4]

[1] Department of Social and Political Sciences, University of Trieste, (e-mail: ddstefano@units.it)

[2] Department of Informatics, University of Salerno, (e-mail: vfuccella@unisa.it)

[3] Department of Economic and Statistics, University of Salerno, (e-mail: mvitale@unisa.it)

[4] Department of Business, Economics, Mathematics and Statistics, University of Trieste, (e-mail: susanna.zaccarin@deams.units.it)

Scientific collaboration is an important driver of research progress that supports researchers in the generation of novel ideas. It has been also recognized as a key factor in measuring and evaluating scientific performance of scholars.

Among the widespread applications of Social Network Analysis (SNA) in the last decades, the study of co-authorship networks, used as a proxy of scholars' collaborative behavior, is one of the topic that most benefited from SNA perspective. Seminal studies explored co-authorship networks in various fields using data gathered from large online international Digital Libraries (DLs) - general (e.g., ISI-WOS, Scopus) or thematic oriented (e.g., Econlit for Economics or Medline for Medical Sciences) - rather than collected by interviews or questionnaires administered directly to the authors of the papers.

Another stream of research focuses on interactions among members of a given target population (e.g., scholars involved in a scientific community or affiliated to a given institution) in order to retrieve the pattern of collaborative behaviors and its effect on the scholars' scientific performance. In this case, recent literature pointed out that international DLs provide a partial coverage of the entire scholar scientific production as well as under coverage of a target population. The integration of international data sources with more specialized and local bibliographic archives can help in the construction of a complete database. Hence in merging different and heterogeneous archives, several issues must be resolved: i) the definition of network boundaries (affecting the type of nodes to be included in the network); ii) the identification of duplicated publication records (affecting network ties); iii) the treatment of scholar synonyms and homonymies (affecting the number of network nodes); and iv) the author name disambiguation of co-authors external to the target population. In this study, we face these issues reconstructing the co-authorship network of a particular scientific community, that is the Italian academic

statisticians. We collect their bibliographic records from an online platform, the Institutional Research Information System (IRIS), available in most of the Italian universities and including international as well as national publications. The platform presents both pros and cons common to other national-based DLs. Even if it guarantees a high coverage rate of our target population and its scientific production, to retrieve co-authorship ties among scholars it is necessary to combine the data contained in different platform deployments available at each university. In addition, data quality is affected by the manual publication data entry made by authors. Moreover, no details are provided on co-authors external to the target population, which implies a huge effort in author name disambiguation.

To deal with these aspects, we first propose a web scraping procedure based on a semi-automatic tool retrieving publication metadata from the online platform in order to reduce the manual adjustments. Second, we introduce a network-based approach to deal with author name disambiguation that requires a minimal set of record attributes (identifier, co-authors, venue). Finally, a discussion on the extension of the proposed procedure in related theoretical contexts will be provided.

KEYWORDS: Bibliographic archives, Co-authorship networks, Web scraping tool, Disambiguation algorithm

SUPERVISED AND UNSUPERVISED HATE SPEECH DETECTION

Emiliano del Gobbo[1], Alice Tontodimamma[1], Lara Fontanella[2] and Luigi Ippoliti[3]

[1] Department of Neurosciences, Imaging and Clinical Sciences, Chieti-Pescara University, Italy, (e-mail: emiliano.delgobbo@unich.it, alice.tontodimamma@unich.it).

[2] Department of Legal and Social Sciences, Chieti-Pescara University, Italy, (e-mail: lfontan@unich.it).

[3] Department of Economics, Chieti-Pescara University, Italy, (e-mail: luigi.ippoliti@unich.it.)

The exponential growth of social media has brought with it an increasing propagation of hate speech and hate based propaganda. Hate speech is commonly defined as any communication that disparages a person or a group on the basis of some characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, religion (Nockleby et al., 2000). Online hate diffusion has now developed into a serious problem: this has led to a number of international initiatives being proposed, aimed at qualifying the problem and developing effective counter-measures. The pursuit of these objectives requires, firstly, the exploitation of statistical techniques for online hate content detection. Classical methods cast the problem in the framework of supervised classification, exploiting algorithms like Naive Bayes, Support Vector Machine, Classification Trees and Neural Networks. However, those techniques are not easily scalable since the task of annotating a very large document collection, as the one that can be easily obtained from Twitter, is strenuous. Therefore, automatic detection techniques have been recently proposed. Those methodologies integrate the Sentiment Analysis (Lin et al., 2012) into an unsupervised classification algorithm, such as the Latent Dirichlet Allocation (LDA; Blei et al., 2003). In our research, we compare the performance of supervised and unsupervised document classification algorithms considering a collection of Facebook comments related to the decision of the President of the Italian Republic to veto the creation of a government after the 2018 Italian general election.

KEYWORDS: Hate Speech, Machine learning, LDA, Sentiment Analyses.

References

Blei D. M., Y. Ng A. Y., Jordan M.I. (2003). *Latent Dirichlet Allocation*. Journal of Machine Learning Research, **3**, 993-1022.

Nockleby John T. (2000). *Hate Speech*. In Encyclopedia of the American Constitution (2nd ed.), L. W. Levy, K. L. Karst et al., New York: Macmillan, pp. 1277-1279.

Lin C., He Y., Everson, R. and Ruger S. (2012). *Weakly-supervised Joint Sentiment-Topic Detection from Text*. IEEE Transactions on Knowledge and Data Engineering (TKDE), **24:6**

THE BALANCE OF INEQUALITY: A REDISCOVERY OF THE GINI'S R CONCENTRATION RATIO AND A NEW INEQUALITY DECOMPOSITION BY POPULATION SUBGROUPS BASED ON A PHYSICAL RATIONALE

Giorgio Di Maio[1], and Paolo Landoni[2]

[1] Department of Economics, Managements and Statistics, University of Milano-Bicocca,
and a/simmetrie – Italian Association for the Study of Economic Asymmetries
(e-mail: giorgio.dimaio@unimib.it)

[2] Department of Management and Production Engineering, Politecnico di Torino,
(e-mail: paolo.landoni@polito.it)

The Balance of Inequality (BOI) is a new approach to the measurement of inequality in which each individual is given a mass equal to his/her income, the population is aligned at regular intervals in order by income, and the center of mass of the income distribution is used to measure its inequality. The BOI index features an intuitive physical interpretation and a simple graphical representation that shows the income distribution and the inequality measure together. When applied to the entire population or a subgroup, the BOI index coincides with the Gini's *R* concentration ratio. The inequality decomposition by population subgroups is obtained for both the *R* concentration ratio (Gini, 1914, 2005) and the Gini index defined from the Lorenz curve with a new approach that considers the distribution of the individual members of each subgroup in the population by introducing the center of mass of the subgroups in the population for cases of perfect equality and perfect inequality, the asymmetry effect and the irregularity effect.

KEYWORDS: Balance of Inequality (BOI) index, Gini index, Inequality decomposition by population subgroups, Inequality measurement, *R* concentration ratio.

References

GINI, C. 1914. Sulla misura della concentrazione e della variabilità dei caratteri. *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti*, **LXXVIII**(Parte Seconda), 1203–1248.

GINI, C. 2005. On the measurement of concentration and variability of characters. *METRON - International Journal of Statistics*, **63**(1), 3–38.

DIMENSIONALITY REDUCTION TECHNIQUES ON THE SIMPLEX FOR TEXT MINING

Simone Di Zio [1], Lara Fontanella[1], Sara Fontanella[2] and Luigi Ippoliti[3]

[1] Department of Legal and Social Sciences, University ”G. d’Annunzio”, Chieti-Pescara, Italy, (e-mail: s.dizio@unich.it,lfontan@unich.it).

[2] Department of Medicine, Imperial College London, London, UK, (e-mail: s.fontanella@imperial.ac.uk).

[3] Department of Economics, University ”G. d’Annunzio”, Chieti-Pescara, Italy, (e-mail: luigi.ippoliti@unich.it.)

From a statistical perspective, text documents can be seen as complex high dimensional objects. As such, they can be represented as vectors in a very high dimensional space, whose dimensions are given by all the terms of interest in the document collection. Dimensionality reduction techniques can help addressing different issues in the text mining context. First of all, reducing data to two or three dimensions facilitates visualisation and working in a lower dimensional space reduces the computational burden of subsequent data analysis. In addition, both document classification and feature extraction can benefit from a lower dimensional representation, since reducing the dimensionality overcomes the curse of dimensionality and, thereby, attenuates the risk of overfitting. In the bag-of-words approach, if the interest lies in relative, and not absolute, differences, the text collection can be represented through a document-term matrix containing the proportions of each term occurrences relative to the document vocabulary. Proportional data are a typical example of compositional data (Aitchison 1982) for which the appropriate sample space is the unit simplex. In this research, we analyse and compare different dimensionality reduction techniques which retain the simplex structure (Kyng et al. 2010; Masoudimansour et al. 2016; Wang et al. 2008).

KEYWORDS: text mining, dimensionality reduction, compositional data, simplex.

References

Aitchison J. (1982). *The statistical analysis of compositional data*, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 44, No.2

Kyng R. J., Phillips J. M. and Venkatasubramanian S. (2010) *JohnsonLindenstrauss dimensionality reduction on the simplex*. 20th Fall Workshop on Computational Geometry, 2010

Masoudimansour W., Bouguila N. (2016) *Generalized dirichlet mixture matching projection for supervised linear dimensionality reduction of proportional data*. 2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP)

Wang H.Y., Yang Q., Qin H. and Zha H. (2008) *Dirichlet Component Analysis: Feature Extraction for Compositional Data*. Proceedings of the 25th International Conference on Machine Learning, 2008

UBU ROI. THE RISE OF ITALIAN POPULISM THROUGH THE ANALYSIS OF SOCIAL MEDIA CONTENT

Ignazio Drudi[1], Fabrizio Alboni[1] and Giorgio Tassinari[1]

[1] Department of Statistical Sciences "P. Fortunati", University of Bologna,
(e-mail: ignazio.drudi@unibo.it; fabrizio.alboni.unibo.it; giorgio.tassinari@unibo.it)

Content analysis has a long tradition in marketing and advertising research, and has been traditionally carried out using ads, initially printed and then on television. Likewise, analysis of literature corpora has a long tradition in linguistic studies. The huge diffusion of social media has widened its range of application. However, besides prototype applications carried out by official statisticians (Istat 2017), the methodology to be used for content and literature analysis of social media data remains largely to be built. In this paper, we aim to contribute to this challenge by introducing and explaining a method to construct a "dictionary (a typical application of text mining techniques) that identifies the meaning attributed to words, based on their use in social media. We then apply this method to study the diffusion of right wing populism on Italian social media, analysing the diffusion of populist "words". To face this empirical task, we refer to the classical definition of populism by Laclau and Mouffe (1985) and by Dornbusch to the more recent researchs (e.g. Chesterley and Roberti (2017)). The results show that the diffusion of populist speech on Italian social media is very similar to the diffusion curve of innovations based on the Volterra-Lotke model. Theoretically, this highlights the relevance of Jarry's argument in Ubu Roi for our understanding of contemporary populist political forces, which underlines the danger of tyranny nested in each protest movement that is not founded on sound theoretical bases

KEYWORDS: populism, Twitter, dictionary, Italy, Volterra-Lotke.

References

CHESTERLEY, N. AND ROBERTI, P. 2017, Populism and Institutional Capture, Bologna: Working Papers DSE n. 1086.

DORNBUSCH, R. & EDWARDS, S. 1991, The Macroeconomics of Populism in Latin America, Chicago: Chicago University Press.

ISTAT, 2017, Big Data Commettee. Annual Report 2017, Roma.

LACLAU, E. & MOUFFE, C. 2001, Hegemony and Socialist Strategy. Towards a Radical Democratic Process, New York: Verso, 2nd Edition

360° UNIVERSITY GOVERNANCE THROUGH BIG DATA

Angela Maria D'Uggento[1], Rosa Ceglie[2] and Massimo Iaquinta[3]

[1] Department of Economics and Finance, University of Bari, Italy
(e-mail: angelamaria.duggento@uniba.it)

[2] Information Technology Unit, [3] Statistical Staff, University of Bari Aldo Moro

Nowadays, it is possible to digitally capture almost all data and store them in high storage capacity devices. The term big data refers to the massive amounts of data from a wide variety of sources, often available in real time. Undoubtedly, big data provide organizations the chance to gain a competitive advantage if data can be analysed effectively to make better business decisions. Lots of Italian researchers are dealing with big data in their studies across a wide range of disciplines, but few Universities seem to have explored the impact that big data management could have on their own organization. The benefits of big data and analytics on higher education institutions are manifold; the paper focuses on the two relevant dimensions of student careers and performance management. The Italian public universities receive state funds partially on the basis of a competitive allocation model in which the number of the high-quality students enrolled and the adoption of a periodic self-evaluation system play a fundamental role. Therefore, universities should use big data to predict academic and behavioural issues of their students, to prevent students from dropping out and, in general, to monitor the predictive variables that lead them to graduation. By means of predictive modelling of data mining, universities could be able to estimate the students preparation, engagement and academic performance at each time of their career. Since 2013 all the Italian Universities have adopted the guidelines about the Self-Assessment, Periodic Evaluation, Accreditation of the degree programmes (AVA) developed by the National Evaluation Agency that is also entrusted with the evaluation of the quality of research activities and of the university collaborations with stakeholders. A second field of application in which big data and analytics can be very useful is the performance management (PM), whose implementation, with the adoption of the performance cycle, highlighted the importance to have an analytics structure available in order to support the governance processes of the organization as a whole. In Italian universities, PM is considered to be the tool to improve efficiency, effectiveness, quality of policies, programs and services. In order to achieve these objectives, the University of Bari has long been developing an "in house" business intelligence system that organizes data flows, inner processes and causal relations. This paper aims at proposing the measurement model adopted by University of Bari as an Italian public university case study to demonstrate its feasibility and the organizational implications when measures are supplied to academic bodies, public sector managers and stakeholders.

KEYWORDS: organizational structures for analytics and big data, university governance, performance measurement.

COMPARING GOAL-BASED AND RESULT-BASED APPROACH IN MODELLING FOOTBALL OUTCOMES.

Leonardo Egidi[1] and Nicola Torelli[1]

[1] University of Trieste,
(e-mail: legidi@units.it, nic.torelli@gmail.com)

There are two distinct approaches on modelling the outcomes of matches in football. The first one involves modelling the numbers of goals scored and conceded between two competing teams in each match directly through a discrete distribution family, e.g. the Poisson distribution.

There are two distinct approaches on modelling the outcomes of matches in football. The first one involves modelling the numbers of goals scored and conceded between two competing teams in each match directly through a discrete distribution family, e.g. the Poisson distribution (Dixon and Coles 1997, Karlis and Ntzoufras 2003, Egidi et al. 2018). A second approach involves modelling win-draw-lose results directly (through multinomial or binomial regression models or their variants, see e.g. Carpita et al. 2015). Clearly, if the focus is the win-draw-lose prediction, the second category is 'nested' within the first one, and gives the finest partition among the three possible results; however, only the first class of models is actually able to estimate the team scoring intensity, and to estimate in a different way team abilities arising from wins of 1-0 rather than wins of 6-1. Including some team-level predictors and fitting the models relying on a historical set of past results are then common features of both the two modelling classes.So far, the choice among these two distinct categories was just a matter of flavor and/or computational burden: we aim at comparing goal-based and result-based models in terms of some probabilistic measures-e.g. Brier score (Brier 1950) and related indexes-and, in general, predictive accuracy. We implement the above comparison both on seasonal national Leagues-e.g. the English Premier League-and the Russia World Cup 2018.Even if the modelling working assumption is strictly Bayesian, we strongly believe our comparison may be broadly applied to the same models considered under a frequentist flavor.

KEYWORDS: Poisson regression, Multinomial logit, Predictive accuracy

References

BRIER, GW (1950) Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78, 1-3

CARPITA M, SANDRI M., SIMONETTO A., ZUCCOLOTTO P. (2015). Discovering the drivers of football match outcomes with data mining. Quality Technology \& Quantitative Management, 12(4), 561-577.

DIXON M. J. and COLES S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. Applied Statistics 46, 265–280.

EGIDI L., PAULI, F., TORELLI, N. (2018). Combining historical data and bookmakers' odds in modelling football scores. Statistical Modelling, 18 (5-6), 1-24.

KARLIS D. and NTZOUFRAS I. (2003). Analysis of sports data using bivariate Poisson models. The Statistician 52, 381–393.

POOLING VIEWPOINTS TO OBTAIN A SINGLE EVALUATION. THE ROI-MOB INDICATOR OF ERASMUS+ MOBILITY EFFECTS

Luigi Fabbris[1] and Manuela Scioni[1]

[1] University of Padua, Department of Statistical Sciences
(e-mail: luigi.fabbris@unipd.it, scioni@stat.unipd.it)

ROI-MOB is the name of an EU project, funded under the Erasmus+ programme, aimed to represent with a single indicator the final outcomes of an international mobility experience. The indicator should include the quality of the experience met by both the young participant, the sending and the hosting units and all other bodies that collaborate to it. The indicator structure is then hierarchical with all possible viewpoints as dimensions to be merged together by means of a set of "importance" weights.

The quality of mobility experiences was measured by the final evaluations of the four stakeholders of each experience, i.e. (i) the participant, (ii) the school or company sending the participant, (iii) the school or company hosting the participant, and (iv) the European Union as general manager of this activity. The weights to attach to the performance measures were estimated through two technical alternatives: (a) the normalised elements of the first eigenvector of the (4x4) non-symmetric matrix of main beneficiaries of international mobility, whose column vectors were defined by the 4 stakeholders answering a specific post-hoc questionnaire, and (b) the normalised elements of the first eigenvector of the skew-symmetric dominance matrix obtained by mediating the stakeholders' dominance matrices constructed with the rankings of international mobility beneficiaries. The results of the comparison between the competing criteria show that both criteria give similar results and, because they imply different computational efforts, allow to take decisions about the computational criterion to adopt in specific circumstances.

KEYWORDS: Composite indicator, Erasmus+ mobility, Aggregation of viewpoints, Weight estimation, Dominance matrix

A PERSONALIZED SMART TOURISM RECOMMENDER SYSTEM BASED ON SOCIAL MEDIA DATA

Daniele Ferone[1], Elisabetta Fersini[1] and Enza Messina[1]

[1]Department of Informatics, Systems and Communication, University of Milano-Bicocca,
(e-mail: daniele.ferone@unimib.it, fersini@disco.unimib.it,
messina@disco.unimib.it)

In the era of Big Data, the information available on the Web may be particularly useful for those users who plan to visit an unknown destination, but the evaluation of a long list of possible points of interest (POI) can be very complex and time-consuming. In this paper, we present a **Tourism Recommender System** (TRS) that using **Machine Learning** techniques is able to *automatically* suggest a ranking list of the N most interesting and relevant POIs to the tourist[1]. The objectives of our TRS are twofold. On the one hand, it provides suggestions to users taking into account their preferences, tastes, and personal interest. On the other hand, the system integrates also information extracted from **social media** to suggest attractions not explicitly targeted on the user, but that are significant in the geographical context. To achieve these goals, the system profiles the users using both explicit and implicit preferences. The explicit ones are provided by the tourists that indicates their main interests, and the implicit preferences are inferred by the system through a **Latent Factor Analysis** (LFA). Applying a **Learning to Rank** (LtR) approach and using both explicit and implicit preferences, the TRS infers which are the most interesting POIs for each user. These results are mediated with a *sentiment score* obtained by performing a **sentiment analysis** on social media data. The score measures the popularity and the appreciation of the attractions located in the geographic area of interest. Choosing the right balance between the two factors allows either to prefer POIs that fit user inclinations or try to suggest significant attractions that are not strictly included in the user preferences but could be of interest for the tourist due to their high popularity in the community.

KEYWORDS: recommender system, smart tourism, machine learning, sentiment analysis.

References

BORRÀS, J. & MORENO, A. & VALLS, A. 2014. Intelligent tourism recommender systems: A survey. *Expert Systems with Applications*, **41(16)**, 7370-7389.

LIU, T. 2009. Learning to Rank for Information Retrieval. *Foundations and Trends in Information Retrieval (FTIR)*, **3(3)**, 225-331.

[1] This work has been partially funded by MISE (Ministero Italiano dello Sviluppo Economico) under the project “SMARTCAL – Smart Tourism in Calabria” (F/050142/01-03/x32).

EXCITABLE TWEETS: SOCIAL COMPUTING AND ONLINE SEXISM

Antonia Anna Ferrante[1], Stamatia Portanova[1]

[1] Technocultures Research Unit Università degli Studi di Napoli L'Orientale

In the last years, a series of racist, sexist, homophobic and transphobic discourses and feelings has been overflowing in all social media. In particular, while a wave of transnational feminism is continuously growing and offering new models of social relations, social media are increasingly becoming the theatres of real attacks against feminist and queer influencers, self-organized groups and information websites. Movements such as Alt right, MRM at a global level, as well as their local Italian nodes, have been greatly investing in social communication strategies, including the use of bots, sentiment analysis, fake accounts, and data analytics. As a result, they have managed to give the impression of being hegemonic in the social communication environment, to the point of launching a series of swarm attacks. In some cases, these attacks have resulted in the closing of accounts and groups, such as the Non Una Di Meno Milano page in 2017.

Data mining and data analytics can certainly act as useful tools for the study of these social phenomena, providing us with a huge amount of information and helping us to find significant patterns, or types of coherent content with particular semantic characteristics, among the collected data. And yet, we also think that, starting from the observation of social data, it is of crucial importance to construct models that can help us to formulate questions and to rethink hegemony online in terms of a counter-narration: are these attacks casual or organised? Is there a way to reconstruct the network behind the observed phenomena? Is it possible to elaborate an inference between social resources and the social matrix of such attacks? In general terms, what are the effects of these 'flames' in terms of safety, as it is perceived by feminist and queer activists? And in what way can these attacks affect the construction of hegemony on the web? This case study might be useful to overcome the usual dicothomy between qualitative/quantitative approach in the social sciences. Accordingly, the analysis of larger patterns of violent social and linguistic behaviour can function in unison with the study of individual details and single case interpretations (a method defined as 'close reading' in the field of Cultural Studies), in order to take the analysis beyond the limits of pure empirical exploration. This method can allow us to undertake a careful study and to devote attention to individual, specific attack cases, in order to arrive to wider, more general reflections about the larger system (or network) behind them.

KEYWORDS: data analytics; online sexism; social media studies; close reading; feminist counternarrative

References

Butler, J., 1997. *Excitable Speech: A Politics of the Performative*, Abingdon-on-Thames: Routledge.

Hall, S., 2006. Notes on Decostructing «the Popular». In *Cultural Theory and Popular Culture: A Reader*. Dorset: Pearson Education Limited, pagg. 477–487.

Lauro, N.C, Amaturo, E., Grssia, M.G., Aragona, B. Marino, M. (eds), 2017. *Data Science and Social Research. Epistemology, Methods, Technology and Applications*. Springer.

Manovich, Lev, 2015. *The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics* (http://manovich.net/index.php/projects/cultural-analytics-social-computing)

PARTICIPANTS' BEHAVIOUR AT SPECIAL EVENTS: SAMPLING PROCEDURES AND GPS TECHNOLOGIES

Mauro Ferrante [1] , Domingo Martín-Martín [2] and Stefano De Cantis [3]

[1] Department of Culture and Society, University of Palermo, (e-mail: mauro.ferrante@unipa.)

[2] Department of Economía Aplicada I, University of Seville, (e-mail: domartin@us.es)

[3] Department of Economics, Business and Statistics, University of Palermo, (e-mail: stefano.decantis@unipa.it)

Events are often used to contribute to the economic development of destinations and to extend tourism seasons. The monitoring of events is particularly useful to evaluate their socio-cultural, economic and environmental impacts. Furthermore, it is helpful for event managers to acquire information on where, when and how activities take place, and the degree of satisfaction about these experiences. Subsequently, the development and implementation of new data collection tools and methods is particularly relevant for the event management (Pettersson & Zillinger, 2011). This paper presents a survey scheme in which traditional survey instruments are used in conjunction with new technologies, such as GPS devices and infra-red beam counter. Moreover, a specific sampling procedure which takes into account for the entry and exit flows of participants at special events is proposed, along with the specification of estimation procedures. The proposed solutions have been implemented on the occasion of the European Researchers' Night 'Sharper 2018' held in Palermo (Italy), a Europe-wide public event dedicated to popular science and fun learning, organized in over 300 cities in Europe and neighboring countries. The single exit/entry point and the relatively brief visiting time, which characterize this and similar events, make the use of GPS technologies particularly suitable for monitoring and evaluation of the event, under the economic, social and environmental perspective.

KEYWORDS: GPS tracking data, tourist mobility, event management, survey methods, European Researchers' Night

References

Pettersson, R., & Zillinger, M. 2011. Time and Space in Event Behaviour: Tracking Visitors by GPS. *Tourism Geographies*, **13**(1), 1–20.

AUTOMATIC MISOGYNY IDENTIFICATION IN ONLINE SOCIAL MEDIA

Elisabetta Fersini[1] and Paolo Rosso[2]

[1] Department of Informatics, Systems and Communication, University of Milano-Bicocca, (e-mail: fersini@disco.unimib.it)

[2] PRHLT Research Center, Universitat Politècnica de València, (e-mail: dprosso@dsic.upv.es)

Hate speech may take different forms in online social media. Most of the investigations in the literature are focused on detecting abusive and/or offensive language in discussions about ethnicity, religion, gender identity and sexual orientation. In this work, we address the problem of automatic detection and categorization of misogynous language in online social media by leveraging on Machine Learning (ML) and Natural Language Processing (NLP) techniques. Since misogyny may take different, we designed a taxonomy to distinguish between misogynous messages and, among the misogynous ones, we characterized the different types of phenomenon and victims. Given the taxonomy, several corpora have been created, by collecting and labeling messages from Twitter in Italian, Spanish and English. Given the corpora, an exploratory investigation on different NLP features (n-grams, pragmatic, syntactic and embedding features) and ML models (Random Forest, Naïve Bayes, Multilayer Perceptron Neural Network and Support Vector Machine) has been performed. By analyzing the results, we can highlight that some linguistic features contribute more on the recognition and classification of misogynistic messages, and that linear models such as Support Vector Machines are a suitable solution for addressing this problem. Finally, we can conclude that the problem of misogyny identification has been satisfactorily addressed, while the misogynistic behavior and victim classification still remains a challenging problem.

KEYWORDS: hate speech, automatic misogyny identification, online social media

References

ANZOVINO, M., FERSINI, E. AND ROSSO, P., 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In I*nternational Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

DIGITAL NEWS PRESS MINING THROUGH TOPIC MODELING, ENTITY RECOGNITION AND SOCIAL NETWORKS ANALYSIS TECHNIQUES

Carlos G. Figuerola[1] and Modesto Escobar[1]

[1]Institute of Studies on Science and Technology, University of Salamanca (Spain), (e-mail: figue@usal.es, modesto@usal.es)

Digital press has became an important media of communication. As individual news can be easily downloaded, they also can be mined in quest of extract valuable information about the social mood in an specific period of time. Besides, topic modeling techniques look for detect and characterize the topics inside a collection of documents (Boyd-Graber et al., 2017). One of the most applied technique is the so called Latent Dirichlet Allocation (LDA). In short, LDA assumes that, in a collection of documents they are a set of topics, each of them in an specific quantity in each of the documents. LDA characterize every topic with a set of words, giving us the proportion of every topic in every document. LDA have a number of implementations, most of them as free software; one of the best known tools is Mallet (http://mallet.cs.umass.edu).

The aim of Named Entity Recognition (NER) is to detect entities ocurring inside a text; entities are things as geographical names, acronyms, enterprises, organizations, personal names, etc. (Nadeau and Sekine, 2007). In this paper we are interested in personal names. NER systems apply machine learning (supervised classification) and deep learning (neural networks) techniques, learning model of contexts of proper nouns. As free software to perform NER we found Polyglot (http://polyglot-nlp.com/) and Spacy (https://spacy.io). Social Network Analysis are well known by sociologists (Scott, 2017). We can use networks to model relationships between persons co-appearing in news. SNA techniques have tools to discover the most influential people, to detect communities of persons or to analyze the evolution of such networks over time. Moreover, we can link persons to the topics of the news.

We have downloaded all the news published by a major spanish newspaper from several years and then we have played around with these tools. The aim of this work is to show some results and discuss the posibilities and potential of such techniques.

KEYWORDS: text mining, topic modeling, entities recognition, social networks analysis.

References

BOYD-GRABER, J., HU, Y., & MIMNO, D. 2017. Applications of topic models. *Foundations and Trends in Information Retrieval*, **11**(2-3), 143-296.

NADEAU, D., & SEKINE, S. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, **30**(1), 3-26.

SCOTT, J. 2017. *Social network analysis*. London: Sage.

SPATIAL LOCALIZATION OF MOBILE PHONE USERS AND TOURISM FLOWS IN SARDINIAN DESTINATIONS' NETWORK

Annamaria Fiori[1] and Ilaria Foroni[1]

[1] Department of Statistics and Quantitative Methods, University of Milano-Bicocca ,
(e-mail: ilaria.foroni@unimib.it, anna.fiori@unimib.it)

Mobile phone providers collect data about the location of their subscribers cellular phones. A mobile phone placing or receiving calls reports its presence to the closest cell towers and communicates its position in the geographical cell covered by one of the towers. Hence, information on the spatial localization of users is contained in the call records of mobile phone carriers. The use of privacy-safe, anonymized datasets represent a huge scientific opportunity to uncover the structure and dynamics of social networks. Quantifying and understanding such patterns may help to obtain deeper insight into applications of great practical importance. For instance, knowing the number of visitors moving from one destination to another is an information that reveals valuable insight for regional tourism planning. In this aim, we consider aggregate data generated by mobile phone users in Sardinia (Italy) to analyze the relationship between networks of tourism destinations and tourism flows. Network analysis is used to draw the patterns of relations among actors and to analyze their structure in the aim of obtaining useful outcomes in the study of tourism destinations. In the first part of our work we study the global network, that is the whole set of destinations and the ways they are linked, while in the second part we investigate more deeply the focal nodes (for instance ports and airports) and the ego networks associated to them.

KEYWORDS: tourism, social network analysis, relational data, spatial connections.

References

LAU, P.L., KOO, T.T., & DWYER, L. 2017. Metrics to measure the geographic characteristics of tourism markets: An integrated approach based on Gini index decomposition. *Tourism Management*, **59**, 171-181.

GONZÁLEZ, M.C., & BARABÁSI, A.L. 2007. Complex networks: from data to models. *Nature Physics*, **3**, 224-225.

SCOTT, N., BAGGIO, R., & COOPER, C. 2008. *Network Analysis and Tourism: from Theory to Practice*. Clevedon: Channel View Publications.

THE CHALLENGES AND LIMITS OF HEALTH OPEN DATA IN ITALY

Carlotta Galeone[1], and Paolo Mariani[2]

[1] Department of Clinical Sciences and Community Health, University of Milan,
(e-mail: carlotta.galeone@unimi.it)

[2] Department of Economics, Quantitative Methods and Business Intelligence, University of Milan – Bicocca
(e-mail: paolo.mariani@unimib.it)

The generation and storage of data have dramatically increased world wide in the last two decades. Computing and networking capabilities combined with openness enhance the potential impact of the accumulated data, offering society an opportunity to drive massive social, political and economic change (Kundra, 2012). Open data is a recent approach. In summary, open data can be freely used, shared and built-on by anyone, anywhere, for any purpose. Though health open data are not regularly available, it is estimated that the value of a more effective use of data resources in the US health care sector alone could be worth USD 300 billion annually (Jetzek, 2015). To date, open government data count more than 10,000 dataset in Italy but only a few concerns healthcare. Other institutional Italian websites (such as Regional open government, AIFA, AGENAS, …) are other important health open data sources, but in general the available healthcare open data are still scarce, with high grade of heterogeneity, not easy to use and link in order to create new useful evidences (Gomes and Soares, 2014). However, a few Italian health care open data projects are on-going with promising results.

KEYWORDS: health care data, Italy, open data, open government data, value generating mechanisms

References

GOMES, Á., & SOARES, D. 2014. Open government data initiatives in Europe: northern versus southern countries analysis. *Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance* (pp. 342-350). ACM

JETZEK, T. 2016. Managing complexity across multiple dimensions of liquid open data: The case of the Danish basic data program. *Government Information Quarterly*, **33**, 89-104.

KUNDRA, V. 2012. Digital fuel of the 21st century: Innovation through open data and the network effect. Joan Shorenstein Center on the Press, Politics and Public Policy

GRASSROOTS-VS- INFLUENCERS CLASSIFICATION IN ANALYSIS OF TWITTER NEWS DIFFUSION

Svitlana Galeshchuk[1,] Ju Qui[2]

[1] Governance Analytics, PSL Université Paris
(e-mail: `svitlana.galeshchuk@dauphine.psl.eu`)

[2] Governance Analytics, PSL Université Paris
(e-mail: `ju.qiu@dauphine.psl.eu`)

This study aims at measuring the relative importance of grassroots in newsmaking and information diffusion on the Twitter social network. It consists of three major methodological steps: (i) classifying Twitter users into grassroots and influencers, (ii) applying topic modelling to detect stories/news in the collection of tweets, (iii) using econometric modelling to analyse and compare the diffusion of stories/news posted by grassroots and influencers. We collected circa 3.5 million tweets and the data on their authors' accounts for the month of May 2018. We developed own Twitter users' classification method using gradient boosting classifier which provides the highest classification accuracy on the test data. Gradient boosting method represents an ensemble of decision trees. Each tree sequentially joins the ensemble correcting the antecedent by fitting its residual errors. Hyperparameters' optimization helped us tune the gradient boosting model. Other machine learning methods deliver worse accuracy on the out-of-sample data: logistic regression, support vector machines, decision tree, random forest. We use the labelled data on 2000 Twitter accounts to train our model with Scikit-learn Python library (see Pedregrosa et al, 2011). We extracted and engineered a number of features which have been used by the classifier to learn how to detect grassroots and influencers. The developed approach is employed to classify the accounts in our May tweets' database. We use LDA topic modelling then. As a next step, we apply a similar regression model as in Banerjee et al (2013) to test the relative importance of the detected grassroots users in the transmission of stories/news on social media to analyse the information diffusion process.

KEYWORDS: social networks, twitter, opinion-making, machine learning, gradient boosting classifier

References

BANERJEE, A., CHANDRASEKHAR, A.G., DUFLO, E. AND JACKSON, M.O., 2013. The diffusion of microfinance. *Scienc.*, **341(6144)**, 1236498.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V. AND VANDERPLAS, J., 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, **12(Oct),** 2825-2830.

Acknowledgment

This work has received support from the research initiative «Governance Analytics » funded by the PSL University under the program «Investissements d'Avenir» launched by the French Government and implemented by ANR with the references ANR-10-IDEX-0001-02 PSL Paris Sciences et Lettres (PSL)

BLOCK-CHAIN ORIENTED SYSTEM FOR THE MANAGEMENT OF PROCESSES

Massimo Giacalone[1], Emilio Massa[2], Diego Carmine Sinitò[2] and Vito Santarcangelo[2]

[1] University of Naples Federico II (e-mail: massimo.giacalone@unina.it)
[2] iInformatica S.r.l.s (e-mail: info@iinformatica.it, sinito.diego@gmail.com, vitho87@hotmail.it)

The following work shows a blockchain-based system for the implementation of control and management of production and supply chain processes. In particular, practical case studies are presented in the context of Industry 4.0 in the field of land safety, car rental, in the production of automotive components and for the dairy production chain conducted in brilliant SMEs in southern Italy. The work also presents the related patents and deliverables implemented at the IT level and also the related approaches for constant monitoring of performances.

KEYWORDS: Blockchain, Industry 4.0, Process Management, Big Data

References

BECK, R., 2017, Blockchain Technology in Business and Information Systems Research, Business & Information Systems Engineering

GIACALONE, CUSATELLI, FANARI, SANTARCANGELO, SINITO', 2018, An innovative approach for the GDPR compliance in Big Data era "Book of short Papers SIS 2018" ISBN 9788891910233, Pearson Editors: Antonino Abbruzzo, Eugenio Brentari, Marcello Chiodi e Davide Piacentino

SANTARCANGELO, V. 2018. Esperienze di Ricerca e Sviluppo applicata alle brillanti realtà del nostro sud, RCE Multimedia.

TASATANATTAKOOL, P., 2018, Blockchain: Challenges and applications, 2018 International Conference on Information Networking (ICOIN)

PRACTICES AND JOURNEYS: INSIGHTS INTO ENVIRONMENTAL ISSUES

Paolo Giardullo, PhD[1]

[1] Department of Philosophy, Sociology, Education & Applied Psychology (FISPPA), University of Padua (e-mail: paolo.giardullo@unipd.it)

The present intervention will contribute to the panel offering a dedicated perspective on data practices (Bates et al 2016) and data journeys (Leonelli 2014) about environmental issues. To talk about sociotechnical construction of data, whatever the scale may be, requires reconsidering how data are constructed. It is crucial to delve into practices that promote data production, use and circulation as materially mediated processes.

Indeed, as Gitelman and Jackson affirm "data produce and are produced by the operations of knowledge production more broadly" (2013 p. 3): this reminds us both of the situatedness of these processes and of their contribution to knowledge production. Indeed, we cannot consider data as pre-existing a specific interest or a set of purposes. Therefore, activities of data collection presume aims, objectives to be achieved. On the other hand, datification is transforming and reshaping several social domains. However, such a reshuffling is not only passively accepted, rather it can be actively performed and bended towards interests other than pure business application. Side to statactivism and data activism, as specific forms of participation as well as contestation by non-institutional actors using data analysis, recently, some environmental issues (e.g. air pollution, oil spills) have been characterised by experiences of participated data collection. To follow practices and journeys of data means to reconstruct trajectories of data, including production, dissemination reconstructing their meaning for actors involved in such processes. These examples about environmental issues give the opportunity to understand sociotechnical processes of data creation offering insights about their political value.

KEYWORDS: sociotechnical construction of data, social practices, environmental issues

References

BATES, J., LIN, Y. W., & GOODALE, P. (2016). DATA JOURNEYS: CAPTURING THE SOCIO-MATERIAL CONSTITUTION OF DATA OBJECTS AND FLOWS. BIG DATA & SOCIETY, 3(2), 2053951716654502.

GITELMAN, L. JACKSON V., INTRODUCTION, IN GITELMAN L., (ED.). (2013). RAW DATA IS AN OXYMORON. MIT PRESS.

LEONELLI, S. (2014). WHAT DIFFERENCE DOES QUANTITY MAKE? ON THE EPISTEMOLOGY OF BIG DATA IN BIOLOGY. BIG DATA & SOCIETY, 1(1), 2053951714534395.

MACHINE LEARNING APPROACHES FOR PRESCRIPTION PATTERNS ANALYTICS

Ilaria Giordani[1], Gaia Arosio[3], Paolo Mariani[2],
Ilaria Battiston[1], Antonio Candelieri[1], Francesco Archetti[1,3]

[1] University of Milano-Bicocca, Department of Computer Science, Systems and Communication
(e-mail: ilaria.giordani@unimib.it, antonio.candelieri@unimib.it)

[2] University of Milano-Bicocca, Department of economics, management and statistics
(e-mail: paolo.mariani@unimib.it)

[3] Consorzio Milano Ricerche, Via Roberto Cozzi 53, 20126 Milano
(e-mail: arosio@milanoricerche.it, archetti@milanoricerche.it)

Abstract. Data analysis processes can give a significant contribution to tackle with the health and socioeconomic threat concerning disease management: exploiting the insights coming from analytical results, health authorities can plan informed actions.

This study summarizes relevant results obtained by data analysis on a dataset with about 650 general practitioners (GPs) and around 1’000’000 patients, during the period 2011 to 2018. With a specific data cleaning process, the sample was reduced to just over 500’000 patients, balanced in terms of gender, age, and therapeutic indication, and related to 140 different types of anti-bacterial agents.

The present study has two objectives: the first part will target the specific antimicrobial resistance problem analysing their anti-bacterial agents prescription dynamics. The insights from this study might allow supporting the transition to an antibiotic resistance surveillance system compliant with the European regulation now being adopted also at the regional level.

The second part of the work focuses on the analysis of the entire database with the aim of the extraction, analysis and evaluation of the Patient Journey, the path performed by each patient during the course of a disease and its treatments. The scenario in which this journey is travelled on is the one of modern healthcare, consisting on complex services that employ several professional figures grouped by professional affiliation, different care settings, and a relationships network based on the dominant organizational system. Over the years, each of these organizational divisions has acquired ever greater autonomy, supporting a specialized care management that does not care of the overall vision. The aim of the proposed analysis is to gain a vision of the disease management, to extract and analyse the data available and to understand which the critical areas are, in order to improve the quality and efficiency of the treatments.

KEYWORDS: patient journey, antibacterial agents, prescription patterns, disease management.

MODELLING HUMAN PREFERENCES BY BAYESIAN OPTIMIZATION

Giordani Ilaria[1], Redivo Attilio[2], Candelieri Antonio[1], Galuzzi Bruno[1], and Francesco Archetti[1]

[1] Department of Computer Science, Systems and Communication, University of Milano-Bicocca,
(e-mail: ilaria.giordani@unimib.it, antonio.candelieri@unimib.it, bruno.galuzzi@unimib.it, francesco.archetti@unimib.it)

[2] OUTCOME s.r.l.,
(e-mail: attilio@outcomeconsulting.it)

In this paper we consider problems in which latent human preferences must be modelled such as in a web design context through A/B testing or recommender systems. These preferences can be queried through pair-wise comparisons only, the so-called duels, so that traditional optimization tools cannot be used, and meta-heuristics are typically adopted. Preference learning has been studied in the context of Gaussian Processes (GPs) which offer a principled way to learn the preference function. We combine an approach based on multi-armed bandits and the capability of the GP to capture, through suitable kernels, correlations across points in the design space. Building on the notion of Copeland score it can be shown that the Condorcet winner is the global optimum of a function whose domain, the design space, can be effectively explored using the tools of Bayesian optimization.

KEYWORDS: Gaussian Processes, Recommender Systems, Bayesian Optimization, Human Preferences

References

KAHNEMAN, D., & TVERSKY, A. 2013. Prospect theory: An analysis of decision under risk. *In Handbook of the fundamentals of financial decision making: Part I*, 99-127.

HERNANDEZ-LOBATO, J. M., HOFFMAN, M. W., & GHAHRAMANI, Z. 2014. Predictive entropy search for efficient global optimization of black-box functions. *In Advances in neural information processing systems*, 918-926.

CANDELIERI, A., & Archetti, F. 2018. Global optimization in machine learning: the design of a predictive analytics application. *Soft Computing*, 1-9.

ARCHETTI, F., & BETRO, B. 1980. Stochastic models and optimization. *Bollettino della Unione Matematica Italiana*, **5** (17-A), 295.

ON THE IMPROVEMENT OF SOCCER MATCH RESULT PREDICTIONS

Silvia Golia [1] and Maurizio Carpita [1]

[1] Department of Economics and Management, University of Brescia,
(e-mail: silvia.golia1@unibs.it, maurizio.carpita@unibs.it)

Many probabilistic and algorithmic models are used to predict the result of a soccer match, that is loss, draw or win of the home team. In order to be useful, this prediction must be produced before the match starts with information available at this time, such as players performance statistics or expert judgements. In general the draw is the most difficult result to forecast (Carpita *et al.* 2018, Strumbelj and Sikonja 2010). This difficulty can be due to the fact that its probability is lower than the probabilities of loss and win, so that the classification models, that use the majority rule as predictive criterion, underestimate the matches resulting in draw. In this study, other predictive criteria are proposed and compared with the modal one, taking into account also the ordinal nature of the result of a match (loss $\prec$ draw $\prec$ win). When the variable is coded as a quantitative one, a rounding criterion must be chosen. In order to evaluate the predictive performance of the proposed criteria, a 3×3 confusion matrix is produces and a set of predictive performance indices built from it, is proposed. The data set derives from the Kaggle European Soccer Database and the variables considered, in addition to the match result, are overall performance indicators for each of the four role in a soccer team, whereas the model used to predict the probability distribution of the match result is the Bayesian Network. Preliminary results highlighted the sensibility of the prediction performances to the choice of the classifier: there are classifiers that are unbalanced towards one of the three results of a match, other more balanced.

KEYWORDS: Polytomous classifier, bayesian networks, performance indices

References

CARPITA, M., CIAVOLINO, E., & PASCA P. 2018. Exploring and modelling team performances of the Kaggle European Soccer Database. To appear in *Statistical Modelling*.

STRUMBELJ, E., & SIKONJA, M.R. 2010. Online bookmakers' odds as forecasts: The case of European soccer leagues. *International Journal of Forecasting*, **26**, 482-488.

GLAM ORGANIZATIONS' DIGITAL MATURITY INDICATOR: A STATISTICAL APPROACH FOR CAMPANIA MUSEUMS

Massimo Guarino [1], Maria Anna Di Palma[1], Antonino Mario Oliviero [2] and Michele Gallo [1]

[1] Dept. of Human and Social Sciences, University of Naples L'Orientale, (e-mail: massimoguarino@outlook.com, madipalma@unior.it, mgallo@unior.it)
[2] Dept of Culture and Society, University of Palermo, (e-mail: antoninomario.oliveri@unipa.it)

Digital Maturity is a complex phenomenon, involving vertical as well as horizontal processes throughout firms. Therefore, the use of multidimensional tools represents the most appropriate approach in measuring companies' digital transformation status (Chanias and Hess, 2016). In recent years, Digital Maturity Models (DMMs)were developed in order to get a current measure of an organization's as-is digital capability (Deloitte, 2018). However, DMMs' effectiveness relies purely on raw data without taking into account items' different difficulties. Moreover, though the vast majority of digital maturity models are mainly focused on manufacturing based organizations, an understanding of the challenges and opportunities identified in each stage of the maturity model can help any organization to identify how digital can lead to success and income generation.The main focus of this work is try to overcome the lack of DMMs applications to the Italian GLAM sector, providing an alternative and more accurate indicator of digital ability , through a widely-used statistical approach. More specifically, we perform a Rasch Analysis using the last national survey on cultural organizations promoted by the Italian Institute of Statistics (ISTAT) in 2016, containing several questions about digital preservation and the use of digital tools in Italian museums.

KEYWORDS: digital museums, cultural heritage, Rasch model

References

Chanias S, Hess T (2016) How digital are we? maturity models for the assessment of a company's status in the digital transformation. Management Report/Institut für Wirtschaftsinformatik und Neue Medien (2):1–14.

Deloitte L (2018) Digital maturity model achieving digital maturity to drivegrowth. Presentation of Deloitte, TM Forum Digital Maturity Model, Feb.

Passing Networks and Game Style in Football Teams: Evidences from European Champions League

Riccardo Ievoli [1], Lucio Palazzo [2] and Giancarlo Ragozini [3]

[1] Department of Statistical Sciences Paolo Fortunati, University of Bologna, (e-mail: riccardo.ievoli2@unibo.it)

[2] Department of Economics and Statistics, University of Salerno, (e-mail: lpalazzo@unisa.it)

[3] Department of Political Sciences, University of Napoli Federico II, (e-mail: giragoz@unina.it)

Summary statistics of football matches give in general poor information about style of play. On such bases, it is generally difficult to quantify how teams are different from each other. This work focuses on the analysis of weighted and directed passing network of football teams. Descriptive measures and structural features of networks are used to evaluate different team strategies by using passage interactions. The main contribution is twofold: on one side is showed how structural properties measured through triadic census are able to distinguish among different styles of play. On the other hand, passing network indices and structural properties are used to better predict probability of winning the match. Data for empirical analysis contain 96 matches in the group stage of UEFA Champions League.

Keywords: network analysis, triad census, correspondence analysis.

A NEW AND UNIFIED THEORY FOR TIME SERIES MODELS WITH ARMA REPRESENTATIONS AND VARYING COEFFICIENTS: ONE SOLUTION FITS ALL

Menelaos Karanasos[1], Alexandros Paraskevopoulos[2] and Alessandra Canepa[3]

[1] Brunel University (UK), (e-mail: menelaos.karanasos@brunel.ac.uk)

[2] University of Pireas (Greece) (e-mail: paraskevopoulosalexandros@gmail.com)

[3] University of Turin (Italy) (e-mail: alessandra.canepa@unito.it)

This research presents a new and unified, in a general sense, theory for time series models with ARMA representations and varying coefficients.

There are two large classes of `time varying' models. The ones with deterministically varying coefficients and those with stochastically varying coefficients. Both type of models have been widely applied in many fields of research, such as economics, finance and engineering, but traditionally they have been examined separately. The new theory unifies them by showing that one solution fits all. We also show that the coherent and complete theory can combine both type of models.

We illustrate mathematically one of the focal points in Hallin's (1986) analysis. Namely, that in a time varying setting two forecasts with identical forecasting horizons, but at different times, have different mean squared errors. This of course, implies that backward asymptotic efficiency (when the initial observation shifts into the remote past) is not equal to forward (termed by Hallin, 1986, Granger-Andersen efficiency) asymptotic efficiency, that is when the time at which a forecast is intended moves into the far future. Since it is meaningless to examine forward asymptotic efficiency in a backward framework, instead we investigate it using a forward model.

Equally important we show how the linear algebra techniques, used to obtain the general solution of the time varying heteroscedastic ARMA model, are equivalent to a simple procedure for manipulating polynomials with time varying coefficients. In order to do so we employ the expression of the Green's function as a Hessenbergian determinant in conjunction with the so called skew multiplication operator or symbolic operator (see, for example, Hallin, 1986, and Mrad and Farad, 2002).

KEYWORDS: ARMA models, time dependent coefficients, forecasting, time varying polynomials

MODELING AND SIMULATING DURATIONS OF A PROFESSIONAL TENNIS MATCH

Francesco Lisi[1] and Matteo Grigoletto[2]

[1] Department of Statistical Sciences, University of Padova,
(e-mail: francesco.lisi@unipd.it)

[2] Department of Statistical Sciences, University of Padova,
(e-mail: matteo.grigoletto@unipd.it)

We develop a model able to describe the duration of professional tennis matches. We explain how to estimate the different "ingredients" needed to compute the duration and we validate our model comparing simulating and observed durations. Finally, we compare durations for different match formats.

KEYWORDS: Tennis analytics, statistics and sport, quantitative analyses in sport.

References

KOVALCHIK, S.A* AND INGRAM, M 2018. Estimating the duration of professional tennis matches for varying formats. *Journal of Quantitative Analysis in Sport*, **14**, 13-23.

THE GLOBAL HEALTH NETWORKS: A COMPARATIVE ANALYSIS OF TUBERCULOSIS, MALARIA AND PNEUMONIA USING SOCIAL MEDIA DATA

Milena Lopreite [1], Michelangelo Puliga [2-3] , Massimo Riccaboni [2]

[1] Dipartimento di Economia, Scuola Superiore Sant'Anna, Pisa (e-mail: m.lopreite@santannapisa.it)

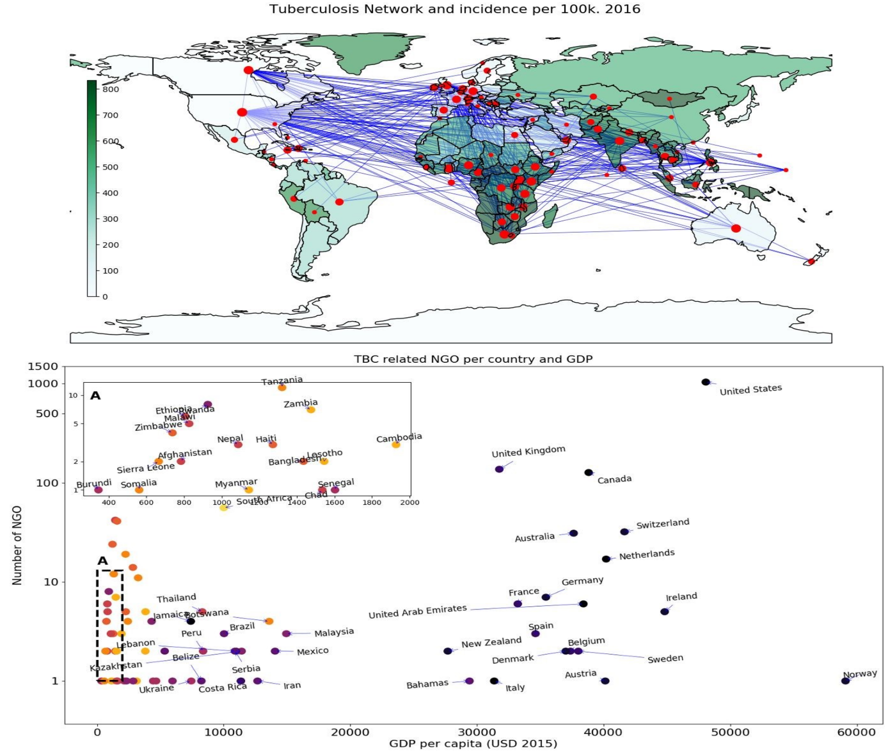
[2] Axis, IMT Alti Studi, Lucca

[3] Linkalab complex networks computational laboratory, Cagliari

Global health networks (GHNs) of organizations fighting major health threats represent a useful strategy to respond to the challenge of mobilizing and coordinating different types of health organizations across borders toward a common goal. In this paper we start from the work of [Shiffman16] and we reconstruct the GHNs of malaria, tuberculosis and pneumonia by creating a new unique database of health organizations collecting data from the official Twitter accounts of each organization. The network is implicit in the friendship relation among each actor, a link connects two organizations that have on Twitter a reciprocal friendship relation. To discover among the Twitter users, the ones that represent organizations or groups active in fighting each disease area we use a Machine Learning classifier (ensemble majority voting with several classifier such as Naive Bayes, and SVM [Bauer1999]). We perform a social network analysis (SNA) of each GHNs to evaluate the structure of the network, the role, and the performance of the organizations in each network (see [Hanneman05]). In particular we measure the network centrality of each node, the connectivity to other links, the popularity in the Twitter platform as ratio of followers over friends, and the total production of tweets. We study also the geographical coverage of each network inferring the location of each organization from its Twitter account. We find evidence that the GHN of TBC, malaria and pneumonia are different in terms of performance, leadership and geographical coverage as well as the Twitter popularity.

The machine learning technique combined with social network analysis may be regarded as a novel tool in the study of global health networks because it suggests key research points in the evaluation of network effectiveness with respect for example to the aid for health as it can be inferred directly from the presence and the actions of the organizations on a social media. Our work is truly interdisciplinary and unconventional in character, by combining computer science, health policy, social sciences with development studies. For example, whether governs and international organizations can be slow or have other more urgent plans, social media are key to rapidly mobilize the international community. They allow a better planning of media campaigns in order to gather new users and to attract more donors for specific goals, eventually mobilizing health policy interventions and global community toward sustainable development goals, especially in low and middle-income countries. A better understanding of the structure and functioning of international social media networks is essential to overcome some of the main challenges in the organization of international aid. Our methodological approach, can be replicated to analyze other international social media networks targeting development goals such as nutrition, poverty, education and environmental protection.

Figure. Top: the global TBC network and TBC incidence 2016, new cases by country per 100000 inhabitants (WHO, 2016).The network is aggregated per country where a block model is superimposed. The size of the node is proportional to the number of organizations in that country. Bottom: the relationship between GPD per capita and number of NGO for every country: the lighter color of the nodes are associated to a higher incidence of the disease.



Keywords: global health policy, social network analysis, machine learning classifier, tuberculosis malaria and pneumonia, health sustainable development.

References

Bauer E. & Kohavi R., An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. 1999. *Machine Learning*, **36**(1):105–139.

Hanneman R.A & Riddle M. 2005 Introduction to social network methods. *University of California online publication*.

Shiffman J., Quissell K., Schmitz H., Pelletier D., Smith S., Berlan D., Gneiting U., Van Slyke D., Rodriguez M., & Walt G.. 2016. A framework on the emergence and effectiveness of global health networks. *Health Policy and Planning*, 31:316.

BASKETBALL SPATIAL PERFORMANCE INDICATORS

Marica Manisera[1], Rodolfo Metulini[1], Marco Sandri[2] and Paola Zuccolotto[1]

[1] BODaI-Lab, University of Brescia,
(e-mail: marica.manisera@unibs.it, rodolfo.metulini@unibs.it, paola.zuccolotto@unibs.it)

[2] DMS StatLab, University of Brescia, (e-mail: sandri.marco@gmail.it)

To assess the scoring probability of teams and players in different areas of a court map is an important topic in basketball analytics, in order to define both game strategies and training programmes.
In this contribution we propose a method based on regression trees, aimed to define a partition of the court in rectangles with maximally different scoring probabilities. Each analysed team/player has its/his own partition, so comparisons can be made among different teams/players. In addition, shooting efficiency measures computed within the rectangles can be used to define spatial scoring performance indicators.

KEYWORDS: basketball analytics, sports statistics, decision trees, performance analysis.

References

ZUCCOLOTTO, P., MANISERA, M., SANDRI, M. 2018. Big data analytics for modeling scoring probability in basketball: the effect of shooting under high-pressure conditions. *International Journal of Sports Science & Coaching*, **13**(4), 569-589.

METULINI, R., MANISERA, M., ZUCCOLOTTO, P. 2018. Modelling the dynamic pattern of surface area in basketball and its effects on team performance. *Journal of Quantitative Analysis in Sports*, **14**(3), 117-130.

PREDICTING CYCLING USAGE FOR IMPROVING BIKE-SHARING SYSTEMS

Giancarlo Manzi[1], Silvia Salini[1] and Cristiano Villa[2]

[1] Department of Economics, Management and Quantitative Methods, University of Milan, Italy, (e-mail: giancarlo.manzi@unimi.it, silvia.salini@unimi.it)

[2] School of Mathematics, Statistics & Actuarial Science, University of Kent, UK, (e-mail: c.villa-88@kent.ac.uk)

Urban mobility is receiving increasing attention as one of the most important dimensions of the so-called *smart city* (Zawieska & Pieriegud, 2018). If mobility must be sustainable, i.e., if it contributes to the improvement of quality of life, bike sharing can be viewed as a possible bridge between wellbeing and economic development. Recent developments in urban management have led bike-sharing systems (BSS) to be a viable complement to traditional public transport systems. However, to understand the mechanisms leading to a successful BSS is a hard task because of the many factors to be considered in its implementation: the docking station network, the number of bikes deployed, the urban morphology, to cite some. One of the most important quandaries is in rightly predicting bike users' behaviour, avoiding an uneven bike distribution among docking stations. In this paper we implement a decision framework to help policy makers to obtain optimal predictions of bike usage in one of the most important BSS in Europe. We use data daily collected from the BSS *BikeMi* in Milan, Italy on each bike itinerary from June 2015 to May 2018, i.e. user and bike ID, check-in and check-out stations, ride timing, bike availability at each station, ride length and check-in and check-out time. We also collected meteorological and environmental indicators like temperature, atmospheric pressure, PM10, NO_2 and CO_2. By using tree-based derived methods (Brieman, 2001) we modelled check out times to detect pick periods of bike usage and rental duration. Results shows that rental duration and weather air pressure were the best predictors for checkout time, whereas checkout time and whether or not the checkout took place in a working day resulted the best predictors for rental duration.

KEYWORDS: bike sharing, transportation, decision trees, random forest, prediction.

References

BREIMAN, L. 2001. Random forests. *Machine Learning*, **45**, 5-32.

ZAWIESKA, J., & PIERIEGUD, J. 2018. Smart city as a tool for sustainable mobility and transport decarbonisation. *Transport Policy*, **63**, 39-50.

GAMING ANALYTICS THROUGH PLAYERS (GAP). PROFILING ITALIAN PLAYERS

Ilaria Mariani[1], Alan D.A. Mattiassi[2] and Emma Zavarrone[3]

[1] Department of Design, Politecnico di Milano (e-mail: ilaria1.mariani@polimi.it)

[2] Department of Economics "Marco Biagi", University of Modena and Reggio Emilia (e-mail: alan.mattiassi@unimore.it)

[3] Dipartimento di Business, Law, Economics and Consumer Behaviour (e-mail: emma.zavarrone@iulm.it)

In recent decades the presence, popularity, and influence of gaming activities in ourdaily lives has progressively increased, making them the object of numerous studies. However, literature research results suggest that the study of playing/gaming habits has mostly been devoted to the videogames area. In consequence, Game Studies have mainly focused on the investigation of digital games, investigating human behavior in relation to this specific medium, rather than exploring the activity itself (Quin et al. 2011; Yee 2006; Seay 2006; Kirby et al. 2014; Desai et al. 2010). Under this perspective a classification approach based on supervised learning algorithms has been applied to data gathered through a survey intended to profile Italian players gaming and social habits when playing, as well as their uses and attitudes, their motivations, frequency and duration of play sessions. In particular, our aim was study and understand Italian players' behaviour, regardless of the medium or device with which the ludic activity is carried out, as well as the specific type of game (Aarseth 2017). To this end, the online survey investigates many forms of play, including digital, live, traditional and role-playing, with an emphasis on the common and transversal aspects of the various types of play, rather than on differences. The questionnaire consisted of a 5-section, 90+ items google form that was disseminated online through social networks and direct email contacts. Over six months, the survey was compiled by 2109 persons. Data provides us several opportunities to vet players as complex systems, from large-scale to more punctual explorations. Here we present data on gaming habits, motivations, social aspects and gender issues. Indeed, the results highlight a taxonomy of gamers in which gender and sociality have a new role.

KEYWORDS: Games, Playing Habits, User Centered Analysis, Players Profiles, Gamers

References

Aarseth, E. 2017. “Just Games.” *Game Studies: The International Journal of Computer Game Research.* 17 (1). Retrieved from: http://gamestudies.org/1701/articles/justgames.

Desai, R., Krishnan-Sarin, S., Cavallo, D., & Potenza, M. (2010). Video-gaming among high school students: health correlates, Gender Differences, and ProblematicGaming. *Pediatrics*, 126, 1414–1424.

Gee, J.P. (2003). What Video Games Have to Teach Us about Learning and Literacy. *Computers in Entertainment* 1 (1): 20.

Griffiths, M.D., Davies, M.N.O., Chappell, D. (2003). Breaking the stereotype: the case of online gaming. *Cyber Psychology and Behavior,* 6(1), 81–91.

Juul, J. (2010). *A casual revolution: Reinventing video games and their players*. MIT press.

Kirby, A., Jones, C., & Copello, A. (2014). The impact of massively multiplayer online role playing games (MMORPGs) on psychological wellbeing and the role of play motivations and problematic use. *International journal of mental health and addiction*, 12(1), 36-51.

Qin, H., Rau, P. L. P., & Gao, S. F. (2011, July). The influence of social experience in online games. In *International Conference on Human-Computer Interaction* (pp. 688-693). Springer, Berlin, Heidelberg.

Seay, A. F. (2006). *Project massive: The social and psychological impact of online gaming*. Carnegie Mellon University, Pittsburgh PA.

Vella, K., Johnson, D., & Hides, L. (2015, October). Playing alone, playing with others: Differences in player experience and indicators of wellbeing. In *Proceedings of the 2015 annual symposium on computer-human interaction in play* (pp. 3-12). ACM.

Vermeulen, L., Van Looy, J., De Grove, F., & Courtois, C. (2011). You are what you play?: A quantitative study into game design preferences across gender and their interaction with gaming habits. In *DiGRA 2011: Think, design, play*. Digital Games Research Association (DiGRA).

Wood, R. T., Griffiths, M. D., & Eatough, V. (2004). Online data collection from video game players: Methodological issues. *Cyber Psychology & Behavior*, 7(5), 511-518.

Yee, N. (2006). Motivations for play in Online games. *Cyber psychology & Behavior,* 9(6), 772–775.

COMPANY REQUIREMENTS AND MONETARY EVALUATION IN THE ITALIAN HEALTHCARE INDUSTRY

Paolo Mariani[1], Andrea Marletta[1], Lucio Masserini[2] and Mariangela Zenga[3]

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, (e-mail: paolo.mariani@unimib.it, andrea.marletta@unimib.it)

[2] Department of Economics and Management, University of Pisa, (e-mail: lucio.masserini@unipi.it)

[3] Department of Statistics and Quantitative Methods, University of Milano-Bicocca, (e-mail: mariangela.zenga@unimib.it)

During last years, many studies have been based on the possibility to give a monetary value for knowledge, skills and attitudes of a candidate during the recruitment process. This work carries out the evaluations of these requirements for individuals that through HR company have placed themselves in the phase of match with the different professional figures. This works will focus the attention on healthcare industry. Starting from data about job vacancies for this sector in Italy in 2017, the aim of the work is to carry out a monetary evaluation of the most important requirements. The analysed requirements have been chosen among a set of soft skills and join with the experience and two knowledge indicators. The methodology used for this work is the choice based conjoint analysis. Using this model, it is possible to identify the features of a candidate that mainly influence the entrepreneurs' choice and the weight of these requirements in the wages. The use of a choice based conjoint model allows to obtain partial utilities that representing the starting point to build a monetary re-valuation index. This index can determine the monetary variation associated with any change in the combination of the attributes of a job with respect to the actual revenue generated by that job.

KEYWORDS: conjoint analysis, labour market, requirements, healthcare industry

References

Lancaster, K.J. (1966). A new approach to consumer theory. In Journal of Political Economy, 74 (2): 132–157.

Mariani, P., Marletta, A., & Zenga, M. (2018). A New Relative Importance Index of Evaluation for Conjoint Analysis: Some Findings for CRM Assistant. Social Indicators Research, 1-14.

THE RISK OF INAPPPROPRIATENESS IN THE ITALIAN GERIATRIC WARDS USING NATIONAL HOSPITAL DATA

Paolo Mariani[1], Andrea Marletta[1], Marcella Mazzoleni[1] and Mariangela Zengai[1]

[1] University of Milano-Bicocca, Italy
(e-mail: paolo.mariani@unimib.it; andrea.marletta@unimib.it; marcella.mazzoleni@unimib.it; mariangela.zenga@unimib.it)

In the last years, the issue of health care for the elderly population is becoming increasingly relevant. Italy could be considered one among the oldest countries in Europe: in fact the population aged 65 and over is 22.6% of the Italian population with an aging index of 168.7% (Istat_a, 2018). Moreover, a high percentage (49.6%) of elderly people shows at least one of chronic/chronic degenerative disease (Istat_b, 2018). This situation, considering an increasing 65-year-old life expectancy, will lead the Italian Health System to cope with a significant increase in healthcare consumption. This work will analyse the ordinary acute admissions in the geriatric wards of the Italian hospitals using the Hospital Discharge Data. Specifically the aim is to identify the risk of inappropriateness of the hospitalizations related to chronic diseases respect to the Italian regions.

KEYWORDS: Italian Geriatric Wards, chronic diseases, Inappropriateness.

References

ISTAT_A (2018). http://dati.istat.it/Index.aspx?DataSetCode=DCIS_INDDEMOG1
ISTAT_B (2018). https://www.istat.it/it/archivio/210557

A MISSING VALUE APPROACH ON FACEBOOK BIG DATA: LIKE, DISLIKE OR NOTHING?

Andrea Marletta [1], Paolo Mariani[1] and Erika Grammatica [2]

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, (e-mail: andrea.marletta@unimib.it, paolo.mariani@unimib.it)

[2] Crea MC, Milan, (e-mail: e.grammatica@campus.unimib.it)

In the last years, the quantity of socio-economic data has grown steeply thanks to the diffusion of Internet and the raise of electronic devices. The Internet diffusion lead to a paradigm change based on new tools. Data extracted from these tools could represent strategic sources to help companies to reach competitive positions for leading the markets. Social media represent an important instrument to mine data and raise the consumers' knowledge. They are able to declare their preferences just giving a "like". This study aims to inspect the mechanism behind users' behaviour. In particular, the attention will be focused on missing expression of the "likes". A statistical model has been proposed to discern a "Like" from a "Dislike" from a "Nothing". The proposed approach could help companies to measure how much a brand is known and how is good to influence users' choices.

KEYWORDS: Big Data, forecasting, Facebook, missing values

References

Little R.J.A. (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association: 1198-1202.

Huisman M. (2009). Imputation of missing network data: some simple procedures. Department of Psychology, University of Groningen, Grote Kruisstraat, the Netherlands.

Rubin D.B. (1976). "Inference and Missing Data." Biometrika n. 63: 581-592.

A Global Rank of the Delphi survey items on the future scenarios of the Family.

Marco Marozzi[1], Mario Bolzan[2]

[1]marco.marozzi@unive.it, Dipartimento di Scienze Ambientali, Informatica e Statistica, Università degli Studi di Venezia.

[2]mario.bolzan@unipd.it. Dipartimento di Scienze Statistiche, Università degli Studi di Padova.

For several decades in many industrialized societies there has been radical, rapid, changing and widespread transformations whose appeal and effects are profoundly affecting the lifestyles and perspectives of our generation. All of this raises questions and reflections that find a natural resonance in the family as they affect their identity and history and those of each of us. In the past quantitative approaches prevailed in the construction of scenarios in particular, today the focus is also on qualitative methods. Nothing can allow us to believe that their evolution will be linear or in any case proportional and symmetrical with respect to the more recent past.

This paper describes some of the results of the research conducted by the Delphi method on the family tomorrow. Some hypotheses of future scenarios useful to understand today and to equip themselves for tomorrow are evaluated by a selected group of experts. From the focus group material, 41 items were identified, divided into 7 sections corresponding to different thematic areas concerning the family. (Parents, spouses, extended family, children, housing, family models, politics and services, communication, solidarity).
The research was carried out in four successive surveys, the first through a face-to-face interview to 32 experts selected according to criteria of expertise on the theme of the family. The subsequent ones using a CAWI method (Computer Assisted Web Based Interview) based on the online self-compilation of a computerized questionnaire, for which the LimeSurvey software was used (www.limesurvey.org). Respondents were asked to provide two assessments. The first "Evolution", or rather the progression of the consistency, of the diffusion of the phenomenon indicated in the item in 10 years, compared to today. The second "Relevance" indicates the importance and added value of the phenomenon expressed by the item. For each item a double evaluation was then requested assigning a value between 0 and 100. After each survey, the evaluations on the evolution and on the relevance expressed by the experts were elaborated in order to submit to the same experts a new questionnaire, characterized by the same questions , but from intervals between I and IX Decile of the distributions of answers provided by the experts in the previous interview. The elaborations presented in the work use the distributions of the last survey. In order to evaluate the relative importance of the various items, the Global Rank approach based on compound indicators was used. This approach allows to evaluate not only the relative importance of the various items in the construction of future scenarios on the family but also the robustness of the results with respect to the particular calculation algorithm used and on which there is no agreement in the literature. An aspect analyzed with great care is that of the different weights to be attributed to the various experts on the basis of their self-assessment about the knowledge of the item.

Key words: Delphi, Composite indicators, Global Rank.

References:

Bolzan Mario, 2018: *Domani in Famiglia*, (A cura di), Franco Angeli Ed., Milano pp: 225.
Marozzi, Marco and Mario Bolzan. 2016. “An index of household accessibility to basic services: A study of Italian regions.” *Social Indicators Research*, DOI 10.1007/s11205-016-1440-0.
Marozzi, Marco. 2015. “Measuring Trust in European Public Institutions”, *Social Indicators Research* 123: 879-895.

SOCIAL MEDIA DISASTERS.
BIG DATA ISSUES IN PUBLIC COMMUNICATION FIELD

Francesco Marrazzo[1] and Gabriella Punziano[1]

[1] Department of Social Sciences, University of Naples "Federico II" (e-mail: francesco.marrazzo2@unina.it, gabriella.punziano@unina.it)

With the growth and changing nature of the (big) data also the role of social sciences researchers has been enhanced, producing an emerging assemblage of tools and techniques for managing and making sense of all this data, often with no more than simple software on a standard computer (Lewis, Zamith, & Hermida, 2013). The future of research on communication field may depend on building intellectual and technical alliances with other ways of knowing (Savage, 2012). The overabundance of data should not end up to be a fake gold (Karpf, 2012), but only much more complicated to analyse (Tinati *et al.*, 2014). Hybrid and mixed solutions (Amaturo & Punziano, 2016) are needed because the structural features of new media can be more fully subjected to algorithmic and quantitative analysis (because of the forms and structures) (Amaturo & Punziano, 2017), while the socio-cultural contexts built up around those features need the careful attention of manual methods and the deepness of qualitative approaches (Marrazzo & Punziano, 2018). In particular, web content analysis (WCA) allows the scholars to expand the horizons of the possible questions that every research can arise in relation to communication and online participation analysis by offering the ability to jointly analyze both the content and the way it is used and re-used in any contexts in which it is realized (Auriemma *et al.*, 2015). WCA hence becomes the basis for the use of techniques that enhance the relational context in which the production of messages and texts puts itself (Amaturo & Punziano, 2013). In the light of these premises, our contribution aims to explore the way in which new research strategies of web content analysis (Herring, 2010) could be useful in the social media disasters implementation process (Rodriguez *et al.*, 2007). As disaster social media framework include users such as communities, government, individuals, organisations, and media outlets (Houston *et al.*, 2015), the use of a broader range of techniques in scientific study of disaster social media effects (Bruns *et al.*, 2012) could facilitate the creation of disaster social media tools in public communication field.

KEYWORDS: Big data, Web Content Analysis, Communities, Social Media Disaster, Public Communication

References

Amaturo, E. and Punziano, G. (a cura di) (2013). *Content analysis tra comunicazione e politica*. Milano: Ledizioni.

Amaturo, E. and Punziano, G. (2016). *I Mixed Methods nella ricerca sociale*, Roma: Carocci.

Amaturo, E. and Punziano, G. (2017). *Blurry Boundaries: Internet, Big-New Data and Mixed-Method Approach*. In Lauro, C., Amaturo, E., Grassia, M.G., Aragona, B. and Marino, M. (2017). *Data Science and Social Research. Epistemology, Methods, Technology and Applications*, New York: Springer International Publishing.

Auriemma, M., Esposito, E., Iadicicco, L., Marrazzo, F., Polimene, A., Punziano, G. and Sarnelli, C. (2015).Euroscetticismo a 5 Stelle: stili comunicativi e online text data nel caso delle elezioni europee 2014. *Sociologia della comunicazione*, 49, pp. 36-54.

Bruns, A., Burgess, J., Crawford, K. & Shaw, F. (2012). *#qldfloods and @QPSMedia: Crisis Communication on Twitter in the 2011 South East Queensland Floods*. ARC Centre of Excellence for Creative Industries and Innovation, Kelvin Grove.

Herring, S. C. (2010). *Web Content Analysis: Expanding the Paradigm*. In Hunsinger, J., Allen, M. and Klastrup, L.(eds.), *The International Handbook of Internet Research*. Berlin and Netherlands: Springer Verlag, pp. 233–249.

Houston, J. B., Hawthorne, J., Perreault, M. F., Park, E. H., Goldstein Hode, M., Halliwell, M. R., Turner McGowen, S.E., Davis, R., Vaid, S., McElderry, J.A. & Griffith, S. A. (2015). Social media and disasters: a functional framework for social media use in disaster planning, response, and research. *Disasters*, 39(1), pp. 1-22.

Karpf, D. (2012). Social science research methods in Internet time. *Information, Communication & Society*, *15*(5), pp. 639–661.

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods. *Journal of Broadcasting & Electronic Media*, *57*(1), pp. 34–52.

Marrazzo, F., Punziano, G. (2018). Online textual data and political communication analysis. Methodological issues and research perspectives. *Sociologia Italiana – AIS Journal of Sociology*, 11, pp. 143-158.

Rodríguez, H., Díaz, W., Santos, J.M. & Aguirre, B.E. (2007). *Communicating risk and uncertainty: science, technology, and disasters at the crossroads*. In Rodríguez, H., Quarantelli, E.L. & Dynes, R.R. (eds.). Handbook of Disaster Research. Springer, New York, NY, pp. 476–488.

Savage, M. (2012). *Identities and Social Change in Britain since 1940: The Politics of Method*. Oxford: Oxford University Press.

Tinati, R., Halford, S., Carr, L., & Pope, C. (2014). Big data: methodological challenges and approaches for sociological analysis. *Sociology*, *48*(4), pp. 663-681.

HUMAN ACTIVITY SPATIO-TEMPORAL INDICATORS USING MOBILE PHONE DATA

Rodolfo Metulini [1] and Maurizio Carpita [1]

In the context of Smart Cities, monitoring the dynamic of the presence of people is a crucial aspect for the well-being of an urban area. We use mobile phone data as a proxy for the total number of people (Carpita & Simonetto 2014), with the specific aim of computing spatio-temporal region specific indicators. Telecom Italia Mobile (TIM), which is the largest operator in Italy, thanks to a research agreement with the Statistical Office of the Municipality of Brescia, provided to us about two years (April 2014 to June 2016) of High-Frequency Daily Mobile Phone Density Profiles (DMPDPs) in the form of a regular grid polygon each 15 minutes. Densities have to be rescaled in order to express the total amount of people rather than just TIM users. Separately for selected regions in the province of Brescia, characterized by being either working or residential areas, we group similar DMPDPs and we characterize groups by their spatial and temporal components. In doing so, we propose a mixed-approach procedure. First, borrowing the method of the Histogram of Oriented Gradients (HOG, Tomasi 2012), we perform a reduction of the DMPDPs dimensionality computing their features extractions. With this method, we convert a 2D spatial object into a 1D vector of data, by preserving the spatial relationship contained in the data. Secondly, we stack in a single vector all the HOG features of the same day and, by applying a high-dimensional cluster analysis that accounts for the *curse of dimensionality*, we group days. Third, for each group, we reshape the data in order to form a 3D array with dimension **a** (*quarters*), **b** (*days*) and **c** (*space*), and we apply a Canonycal Polyadic (CP) tensor decomposition (CANDECOMP/PARAFAC, Kolda & Bader 2009) to extract three indicators related to the dynamic of the presences along the space, the days and the quarters.

KEYWORDS: image clustering, big data, urban planning, spatio-temporal indicators.

References

Carpita, M., Simonetto, A. (2014). Big Data to Monitor Big Social Events: Analysing the mobile phone signals in the Brescia Smart City.*Electronic Journal of Applied Statistical Analysis: Decision Support Systems*, **5(1)**,31-41.

Metulini, R., Carpita, M (2018). On Clustering Daily Mobile Phone Density Profiles. *High Dimensional Small Data Workshop*, Ca Foscari (Venice).

Tomasi, C. (2012). Histograms of oriented gradients. *Computer Vision Sampler*, 1-6.

Kolda, T.G., Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, **51(3)**, 455-500.

[1] Data Methods and Systems Statistical Laboratory, Department of Economics and Management, University of Brescia (e-mail: rodolfo.metulini@unibs.it, maurizio.carpita@unibs.it)

THE INFLUECE OF CITIES ON INTERGENERATIONAL SOCIAL MOBILITY

Alessandra Michelangeli[1], Umut Türk[2]

[1] Department of Economics, Statistics and Marketing, University of Milano-Bicocca, (e-mail: alessandra.michelangeli@unimib.it)

[2] Department of Economics, Abdullah Gül University, Sumer Campus, Kayseri (Turkey), (e-mail: umut.turk@agu.edu.tr)

In this paper, we develop a theoretical framework about intergenerational social mobility based on a wider concept of upward mobility. The latter usually refers to changes in social status between different generations within the same family. Individuals may improve their socioeconomic status thanks to a series of factors or events, such as education, job changes, career advancements. We focus on the role played by higher education and we also consider, as additional factor, the fact of moving to another city to attend university. Cities to which students move may offer resources and opportunities that amplify the human capital accumulation through education.

We also provide an empirical application to determine the influence of Italian cities in favouring upward mobility of graduated people who moved to another city to get higher education. The empirical strategy fully exploits the spatial dimension of student migration flows.

KEYWORDS: intergenerational social mobility, city, spatial mobility.

A SUPERVISED LEARNING APPROACH IN THE RISK ESTIMATE OF GAMBLING IN ADOLESCENTS: A CASE STUDY

G S Monti [1], L Benedan[2] and M Mercandelli [3]

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, (e-mail: gianna.monti@unimib.it)

[2] Bicocca Applied Statistics Center, University of Milano-Bicocca (e-mail: laura.benedan@unimib.it)

[3] e-mail: m.mercandelli5@outlook.it

Adolescent gambling is internationally considered a serious public health concern, although this phenomenon is less explored with respect to adult gambling. It is also well known that the early onset age of gambling is a risk factor for developing gambling problems in adulthood (Dowling et al., 2017).

This study examined 9671 adolescents enrolled in 16 public high schools in Lombardy, a northwest Italy region, between March 2017- April 2018 and it is part of a larger study aimed at investigating dysfunctional behaviours of adolescents, with the purpose of identifying the factors that increase the risk of vulnerability and the protection factors able to reduce the incidence of pathological phenomena. The objective of the present study was to investigate the vulnerability of adolescents in high school to develop gambling problems, explained by some individual and social factors, and by the association with the use of substances, such as alcohol and tobacco, and other risk-taking behaviours. A penalized logistic regression analysis was used (Hastie et al., 2001) and various machine learning methods were compared to deal with imbalanced classes, as the prevalence rate of high school students in the sample with problem gambling is equal to 5%.

KEYWORDS: gambling, class imbalance, machine learning.

References

Dowling, N. A., , Merkouris, S. S., Greenwood, C. J., Oldenhof, E., Toumbourou, J. W., and Youssef, G. J. (2017). Early risk and protective factors for problem gambling: A systematic review and meta-analysis of longitudinal studies. *Clinical Psychology Review*, 51:109 – 124.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

REGULARIZED SEMIPARAMETRIC ESTIMATION OF HIGH DIMENSIONAL DYNAMIC CONDITIONAL COVARIANCE MATRICES

Claudio Morana[1]

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, Center for European Studies (CefES), Center for Research on Pensions and Welfare Policies (CeRP) and Rimini Centre for Economic Analysis (RCEA). (e-mail: claudio.morana@unimib.it)

In this paper we address the issue of consistent estimation of time-varying conditional correlations in high dimensional settings, within the context of the Semiparametric DCC model proposed by Morana (2015). While sharing a similar sequential approach to DCC (Engle, 2002), SP-DCC has the advantage of not requiring the direct parameterization of the conditional covariance or correlation processes, therefore also avoiding any assumption on their long-run target. It can be implemented in a high dimensional context, without downward biased estimation, since it only requires univariate QML GARCH estimation for individual and pairwise aggregated series in the first step; conditional covariances are then estimated in the second step by means of the polarization identity. Monte Carlo results, reported in Morana and Sbrana (2017), yield support to SP-DCC estimation in various parametric settings of empirical interests for financial applications. Relative to the two-step SP-DCC model, this paper contributes an additional, third step, which entails ex-post regularization of the conditional covariance and correlation matrices and an efficiency gain. In this respect, a new regression-based non-linear shrinkage approach is proposed, which ensures accurate estimation and safe inversion of the conditional covariance and correlation matrices also in high dimensional cases. Moreover, optimal smoothing of the conditional correlation and covariance matrices, grounded on the maximization of the joint likelihood of the model, is performed. Relative to available DCC approaches, in addition to efficiency improvements, the new regularized SP-DCC (RSP-DCC) is expected to grant higher estimation accuracy, as it allows for more flexible modelling of second moments and spillover effects of past conditional variances and innovations on the conditional covariances. We apply the proposed approach for the estimation of a global minimum variance portfolio using fifty assets. The empirical results confirm that SP-DCC is a simple and viable alternative to existing DCC models.

KEYWORDS: conditional covariance, dynamic conditional correlation model, semiparametric dynamic conditional correlation model, multivariate GARCH.

References

ENGLE, R.F. 2002. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models, *Journal of Business and Economic Statistics*, **20**, 339-350.

MORANA, C. 2015. Semiparametric Estimation of Multivariate GARCH Models, *Open Journal of Statistics*, **5**, 852-858.

MORANA, C., & SBRANA, G. 2017. Some Financial Implications of Global Warming. An empirical assessment, University. of Milan Bicocca DEMS WP No. 377, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3092970.

SMART WORK AS AN EMPLOYEE WELFARE PRACTICE: AN EXPLORATIVE RESEARCH PROJECT

Ariela Mortara

[1] Department of Business, Law, Economics, and Consumer Behaviour, Università IULM di Milano (e-mail: ariela.mortara@iulm.it)

In the last years, the labour market has undergone many changes and the label Work 4.0 is often used to highline new prospects and opportunities for shaping further developments. Work 4.0 refers to employment conditions in the context of an extensive shift in values such as standard employment relationships, individual flexibility in developing life plans, and social protection introducing changes in corporate culture. Big data, Digital transformation, Crowdworking emerge as features deeply influencing the labour market (Salimi, 2015).

In Italy, the legislation has introduced through law number 81/2017 smart working as an innovative tool instead of telework. Indeed, Italian companies identified the smart working paradigm with communication solutions and devices for working on and off the company premises. However, smart working guarantees to the worker the possibility of reconciling life and work times (Kossek and Ollier-Malaterre 2013; Fleetwood 2007) and, as for the employer, it allows better work organization in terms of productivity and cost reduction. Thus, smart working, as a tool to facilitate work-life balance, falls into the employee welfare practices.

The paper presents some results of a quantitative research project aimed at understanding the importance that workers attribute to the various initiatives of employee welfare according to different intervention areas.

KEYWORDS: employee welfare, smart work, work-life balance

References

FLEETWOOD, S. 2007. Why work-life balance now? *The International Journal of Human Resource Management*, 18(3), 387-400.

KOSSEK, E. E. & OLLIER-MALATERRE, A. 2013. Work-family policies: Linking national contexts, organizational practice and people for multi-level change. In Poelmans S., Greenhaus J., Las Heras Maestro M. (eds.). *New frontiers in work-family research: A vision for the future in a global world*. Basingstoke, England: Palgrave Macmillan, pp. 1–53

SALIMI M. 2015. WORK 4.0: An Enormous Potential for Economic Growth in Germany, *@Adapt_bulletin*, retrievable at http://englishbulletin.adapt.it/wp-content/uploads/2015/12/work4.0_sam.pdf (26 November 2018).

VIRTUAL ENCOUNTER-SIMULATIONS: A NEW METHODOLOGY FOR GENERATING CONFLICT DATA

Georg P. Mueller [1]

[1] Faculty of Economics and Social Science, University of Fribourg (Switzerland)
(e-mail: Georg.Mueller_Unifr@bluewin.ch)

Interview data are generally "monadic": they report about properties, attitudes, and values of particular interviewed persons but rarely about their relations to others like e.g. interpersonal conflicts. For this reason the author proposed in two earlier publications (Mueller 2011, 2017) to construct *dyadic* data records by combining the information of pairs of *randomly* selected persons. By aggregating the value differences of these pairs it is possible to explore conflicts not only *between* but also *within* groups. The latter intra-group conflicts are insofar important as they serve as *benchmarks* for the evaluation of inter-group conflicts: it is assumed that an inter-group conflict is only salient if it is higher than an associated intra-group conflict. Since there are *two* group-specific conflict-benchmarks the same inter-group dissent may be salient from the perspective of one of the groups but *not* for the other. Consequently, virtual encounter simulations are suitable to identify *asymmetrical* conflicts, which is not possible with traditional statistical methods based on mean values of groups. Moreover, by the use of intra-group conflicts as benchmarks, a certain amount of inter-group conflict is "normal". Consequently, with the virtual encounter method small value differences between groups may be insignificant, contrary to the results of traditional t-tests of the related group specific mean values.

For illustrative purposes the proposed virtual encounter simulations are applied to the political east-west conflict between German- and French-speaking Switzerland, which often leads to different regional outcomes in national plebiscites (Milic et al. 2014: 191 ff.). Based on interview data of the ISSP (2013), the empirical analysis of the paper attempts to answer the question, whether the two language groups have different *national identities*. The virtual encounter-simulations show that German-speaking Switzerland is relatively homogeneous with regard to national identity and consequently has *asymmetrical* conflicts with the politically less homogeneous French speakers. There are however many aspects of national identity, where the virtual encounters method finds *no conflict* between Eastern- and Western-Switzerland. These findings are in a final discussion compared with the results of traditional t-tests of mean political attitudes, which generally show stronger but no asymmetrical conflicts.

KEYWORDS: Simulation, virtual encounters, conflict, interview data, national identities.

References

ISSP 2013. National Identity III. *https://www.gesis.org/issp/modules/issp-modules-by-topic/national-identity/2013/* (edited by GESIS).

MILIC et al. 2014. *Handbuch der Abstimmungsforschung (Handbook of research about plebiscites)*. Zurich: Verlag Neue Zürcher Zeitung.

MUELLER, G. 2011. Microsimulation of Virtual Encounters: A New Methodology for the Analysis of Socio-Cultural Cleavages. *International Journal of Microsimulation* **4**, 21-34.

MUELLER, G. 2017. On the Use of Microsimulation for Investigating Ideological Dissent: Exemplary Analyses of the Values of the European Political Left. *ASK* **26**, 61-80.

FROM RECOGNITION TO RE-USE: OPEN DATA FOR CONFISCATED GOODS

Giuseppe Notarstefano[1], Umberto Di Maggio [1] and Giuseppe Ragusa[2]

[1] Department of Law, LUMSA of Palermo, (e-mail: g.notarstefano@lumsa.it, u.dimaggio@lumsa.it)
[2] ON-DATA group, (e-mail: giuragu@gmail.com)

Reuse confiscated goods (also applies to firms) has a duple civic meaning: primarily that of a social return of Mafia assets, but also a recognition of a new category of goods, common goods precisely. This paper intends to propose an analysis of the information base on the confiscated goods and the growing need for open data to improve the transparency of management processes.

The main source is provided by the National Agency for the administration and destination of assets seized and confiscated from organized crime (ANBSC), while at the European level the number derives from the analysis of stats produced by the Assets Recovery Offices (ARO) or by national Ministries of Interior and Justice. The main purpose is clearly represented by the definition of a reliable and "real" mapping of the actual endowment of the assets. But to achieve this, it is considered essential to put on a more participated and fed bottom up system.

The "Confiscati Bene 2.0" project pursues precisely this goal: emphasis on Open data is essential: the partnership between institutions and associations promote a culture of transparency and provides closer monitoring to actual conditions of such goods. It is not just a matter of technical or methodological aspects, but of a concrete social recognition that assigns a leading role to civil society. Transparency also becomes a non-accessory requirement even in the pursuit of the will of the Law, the widespread monitoring that it is able to activate reinforces the law enforcement action that the judicial authority plays against organized crime.

The goal is also to enrich the statistical information, adding to the database as well as an extensive meta information also information on best practices.

KEYWORDS: confiscated goods, open data, Confiscati Bene 2.0

References

DI MAGGIO, U., NOTARSTEFANO G. & RAGUSA, G. 2018. Ri–conoscere i beni confiscati Un percorso tra partecipazione, condivisione e trasparenza in INGRASSIA, R. (EDS) *Economia, organizzazioni criminali e corruzione*, Aracne editore 161-178.

http://ondata.it/

http://eu.confiscatibene.it/

PROMOTING CRUISE SHIP AS "TOURIST DESTINATION" ON TELEVISION: THE CASE OF ITALY

Antonino Mario Oliveri[1] and Gabriella Polizzi[2]

[1] Department Cultures and Societies, University of Palermo,
(e-mail: antoninomario.oliveri@unipa.it)

[2] Faculty of Human and Social Sciences, University of Enna "Kore" (e-mail: gabriella.polizzi@unilore.it)

KEYWORDS: destination image, cruise ship, TV, textual analysis.

Destination image is one of the elements that most affect tourists' decision-making processes (Baloglu and McCleary, 1999). Since a distinctive image can differentiate a destination from its competitors, destinations usually compete also via images (Urry, 1990). In this regard, cruise tourism is no exception, since "image is what sells cruises" (Klein, 2002). Over the past 30 years this has led to significant growth in advertising activities aiming to promote the cruise ship "as the destination in itself" (Wood, 2004). Similar trends can be observed among Italian cruise lines as well. Among different types of promotional texts, the importance of the visual ones in shaping the ways tourist represent, choose, consume and recall a place (Berger, 1972; Urry, 1990) has also been acknowledged in the field of cruise tourism.
Starting from this background, this paper presents the methodology and main results of textual analysis carried on a sample of recent TV commercials broadcast by the major cruise lines operating in Italy. The final aim of this research is to identify what attributes of cruise ship as tourist destination have been portrayed, as well as what specific profiles of Italian cruisers and consumption styles have been advertised.

References

BALOGLU, S., & MCCLEARY, K.W. 1999. A model of destination image formation. *Annals of Tourism Research*, **26**(4), 868–897.

BERGER, J. 1972. *Ways of Seeing*. London: Penguin Books.

KLEIN, R.A. 2002. *Cruise Ship Blues: The Underside of the Cruise Industry*. Gabriola Island, British Columbia, CAN: New Society Publishers.

URRY, J. 1990. *The Tourist Gaze. Leisure and Travel in Contemporary Societies*. London: Sage.

WOOD, R.E. 2004. *Cruise Ships: Deterritorialized Destinations*. In: LUMSDON, L. and PAGE, S.J. (Eds.), *Tourism and Transport: Issues and Agenda for the New Millennium* (pp. 133–145). Amsterdam: Elsevier.

MEASURING TOURIST SATISFACTION AND DISSATISFACTION: ADAPTATION OF THE 4Q METHODOLOGY TO THE CASE OF WEB BASED DATA

Antonino Mario Oliveri[1], Gabriella Polizzi[2], Anna Maria Parroco[3] and Michele Gallo[4]

[1] Dipartimento Culture e Società, Università degli Studi di Palermo (e-mail: antoninomario.oliveri@unipa.it)

[2] Facoltà di Scienze dell'Uomo e della Società, Università degli Studi di Enna "Kore" (e-mail: gabriella.polizzi@unilore.it)

[3] Dipartimento di Scienze Psicologiche, Pedagogiche, dell'Esercizio Fisico e della Formazione, Università degli Studi di Palermo (e-mail: annamaria.parroco@unipa.it)

[4] Dipartimento di Scienze Umane e Sociali, Università degli Studi di Napoli L'Orientale (e-mail: mgallo@unior.it)

KEYWORDS: tourist satisfaction and dissatisfaction, 4Q methodology, web based data.

Tourist satisfaction has been studied so far using many different theoretical approaches and measurement models. In doing that, scholars have considered satisfaction and dissatisfaction either as two extremes on a single continuum or two distinct conceptual dimensions, i.e. tourist satisfaction (TS) and tourist dissatisfaction (TD) (Alegre and Garau, 2010). In line with the latter approach, known as *dual approach*, the 4Q is a methodology which permits to measure TS and TD through the administration of just four open-ended questions to tourists (Oliveri et al., 2018). The main feature of the 4Q methodology is that data regarding dozens of TS and TD elementary indicators can be drawn from the answers provided by tourists to the four questions. Afterwards, composite indicators of both TS and TD can be constructed. This work aims to propose some adaptations of the 4Q methodology to the analysis of TS and TD as emerged from spontaneous online travel reviews, and to show how to obtain TS and TD composite indicators which are able to meet the recommendations issued by OECD-JRS (2008).

References

ALEGRE, J., & GARAU, J. 2010. Tourist satisfaction and dissatisfaction. *Annals of Tourism Research*, **37**(1), 52–73.

OECD & JRC 2008. *Handbook on constructing composite indicators. Methodology and user guide*. Paris: OECD.

OLIVERI, A.M., POLIZZI, G. & PARROCO A.M. 2018. Measuring Tourist Satisfaction Through a Dual Approach: The 4Q Methodology. *Social Indicators Research*, 1–22.

DETECTING MULTIDIMENSIONAL CLUSTERING ACROSS EU REGIONS

Pasquale Pavone[1], Margherita Russo[1], Francesco Pagliacci[1], Simone Righi[2] and Anna Giorgi[3]

[1] Dipartimento di Economia Marco Biagi, and CAPP, UniMORE, Italy (e-mail: pasquale.pavone@unimore.it, margherita.russo@unimore.it, francesco.pagliacci@unimore.it,)

[2] Department of Computer Science, UCL, United Kingdom, (e-mail: s.righi@ucl.ac.uk)

[3] Leader AG1 EUSALP Lombardy Region representative, and Gesdimont research centre, University of Milan, Milano, Italy (e-mail: anna.giorgi@unimi.it)

This paper applies multidimensional clustering of EU-regions to identify similar specialization strategies. Different techniques are applied to an original dataset, created by the research team, where EU-28 regions are classified according to their socioeconomic features and to the strategic features of their research and innovation smart specializations strategy (RIS3). In the first classification, each region is associated to one categorical variable (with 19 modalities). In the classification of RIS3, two clustering of "descriptions" and "codes" of RIS3 priorities were considered (respectively made of 23 and 21 Boolean categories).

The configuration of data is discussed in the paper. Two techniques of clustering have been applied: Correspondence Analysis and Infomap multilayer algorithm. The most effective clustering, in terms of both the characteristics of the data and the emerging results, is the one obtained with a Correspondence Analysis. On the contrary, given the very dense network does not produce significant results when Infomap is applied. A classification of regions encompassing the three dimensions under analysis is of particular interest in the current debate on post 2020 European Cohesion Policy, aiming at orienting public policies on the reduction of regional disparities.

KEYWORDS: smart research and innovation strategies, multi-dimensional analysis, clustering, European regions

References

RUSSO M., PAGLIACCI F., PAVONE P., GIORGI A. 2018
RIS3 IN MACROREGIONAL STRATEGIES: TOOLS TO DESIGN AND MONITOR INTEGRATED TERRITORIAL DEVELOPMENT PATHS. 4_ESPON SCIENTIFIC CONFERENCE 2018 LONDON, UNITED KINGDOM, 14TH NOVEMBER 2018

ESTIMATING HIGH DIMENSIONAL STOCHASTIC VOLATILITY MODELS

Matteo Pelagatti[1] and Giacomo Sbrana [2]

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, (e-mail: matteo.pelagatti@unimib.it)

[2] Department of Information Systems, Supply Chain & Decision Making, NEOMA Business School (e-mail: Giacomo.SBRANA@neoma-bs.fr)

Models for time series with time-varying variances and covariances have become very popular in finance because they capture features that are typical of many asset returns, such as volatility clustering and heavy tails. Two main econometric approaches emerged in the literature: ARCH-type models and stochastic volatility (SV) models. ARCH-type models have the advantage of being simple to estimate, while SV models are harder to implement but are backed by financial theory.

For a long period of time, the multivariate versions of both models have been only moderately successful since their estimation on realistically large portfolios of assets was not feasible. For ARCH-type models a convincing solution to the curse of dimensionality was proposed by Robert Engle with his Dynamic Conditional Correlation (DCC) model. No equally successful model has been introduced in the SV world.

In this work we derive three general results that can be jointly used to estimate high dimensional multivariate SV models cast in linear state-space form such as the one in Harvey, Ruiz and Shephard (1994), from now on HRS, and a multivariate extension of the one in Alizadeh, Brandt and Diebold (2002), from now on ABD. Let us call *d* the number of time series and *n* their (common) length. The problem with the approaches of HRS and ABD is that they cast their models in state-space form and carry out Gaussian quasi maximum-likelihood (QML) estimation using the Kalman filter and numerical optimisation. Each pass of the Kalman filter implies, for every time point $t \in \{1,2,...,n\},$ sums, multiplications and an inversion of $d \times d$ matrices. Thus, for large *d* the computational burden becomes too expensive. Furthermore, the typical quasi-Newton optimisers used to maximise the log-likelihood function become very unstable when the number of parameters is very large.

Our solution to the aforementioned issues in estimating large HRS and ABD models is based on three results. Firstly, we substitute the Kalman filter recursions with the steady-state Kalman filter, which we obtain in closed form for the multivariate AR(1) plus noise model. This approximation does not harm the asymptotic properties of the parameter estimates and reduces by a factor of *n* the number of operations on the $d \times d$ matrices of the regular Kalman filter. Secondly, we design an EM algorithm based on the steady-state filter and smoother to be used in substitution of quasi-Newton optimisers. Our algorithm is numerically very stable and, as any EM algorithm, it moves very quickly towards a neighbourhood of the solution. Finally, we propose a simple estimator of the correlation between returns that is consistent and asymptotically normal regardless of the evolution of the variances, provided that the returns are drawn from elliptical distributions with time invariant correlations. This last result allows the quick estimation of one of the two large covariance matrices of the multivariate SV models, which, as it will be clear later, is only a deterministic transformation of the correlation matrix of the returns.

KEYWORDS: Riccati equation, EM algorithm, Kalman filter, Large covariance matrix, Multivariate stochastic volatility

EXPLORING PATHS THROUGH PLACETELLERS PERFORMATIVITY ON INSTAGRAM

Ilaria Primerano[1], Giuseppe Giordano[1]and Pierluigi Vitale[1]

[1] Università degli studi di Salerno, (e-mail: iprimerano@unisa.it, ggiordano@unisa.it, pvitale@unisa.it)

In the last decades, the spread of Internet and online social media has created a huge amount of data that is able to provide new insights to researchers in different disciplinary fields. Much of these data can be easily coded as relational data. In this study we apply an interdisciplinary approach based on Visual Content Analysis, Social Network Analysis and explorative statistical techniques. Specifically, we use data extracted from an online social network (i.e. Instagram) to identify travellers' paths among sites of interests. We select the most important cities in the Campania region linking them according to the common geolocalization of the published images in a given time frame. Starting from a huge collection of all the pictures labelled with different cities' names, we define three training set in order to obtain three different evocative categories through placetellers performativity, such as food, person portrait, places, and so on. An expert and supervised system is trained to recognize the context and such information is used to illustrate the emerging paths. The data collection can be organized and described through a network on which the main paths will be analysed. Statistical network centrality measures are used to identify relevant places and their characterizations.

KEYWORDS: data visualization, machine learning, social media, social network analysis, visual content analysis

DIGITAL NATIVES BUT NOT YET DIGITAL CITIZENS: HOW THE DIGITAL GAP AFFECTS THE EDUCATIONAL POVERTY OF YOUNG PEOPLE

Quattrociocchi Luciana [1] and Grassia Gabriella [2]

Educational poverty is a multidimensional phenomenon. In Italy it is considered as the result of the many inequalities due to the difficulty in the socio-economic conditions and the territorial context in which we live and grow and which have an impact on the cognitive development of the youngest.

Istat has the task of identifying territorial indicators of educational poverty for policy interventions aimed at combating educational poverty.

In our opinion, effect of all the existing cultural divides the digital divide, for too long underestimated and little depth, represents a new aspect of educational poverty that further distances the most vulnerable young people from their peers. While the vast majority of 15-29 year-olds have acquired a certain familiarity with digital tools, which crosses all aspects of their personal, family and school life, there is a significant minority of *digital native* who are excluded from a precious resource for its growth. Raising young people from the risk of educational poverty therefore means granting them access to knowledge that today can not be separated from digital literacy (Literacy in a digital world).

In this work, our aims are:

- to illustrate the progress of research in ISTAT on educational poverty,
- to explore how many young people are still excluded from access to new digital technologies,
- to examine the reasons for digital exclusion,
- to identify what types of literacy people need to obtain in an increasingly digital and interconnected society.

KEYWORDS: "multidimensional educational poverty, composite indicators, official statistics."

References

Mazziotta, M., and Pareto, A. (2016). "On a generalized non-compensatory composite index for measuring socio-economic phenomena." Social indicators research, 127(3), 983-1003

Save the Children (2016) "Liberare I bambini dalla povertà educativa. A che punto siamo?" available at url: https://www.savethechildren.it

[1] ISTAT, (e-mail: luciana.quattrociocchi@istat.it)

[2] University Federico II, Naples

“WORKING SMART” IN A DIGITAL CONTEXT: FROM DIGITAL TECHNOLOGY TO DIGITAL COMPETENCES

Aurelio Ravarini

[1] Università Carlo Cattaneo (e-mail: aravarini@liuc.it)

Digital technology enabled phenomena (such as IoT, big data, social media) are radically transforming several industries and is making many jobs obsolete, while we see a growing demand of new types of skills, beyond the technical ones (Wang 2012, Strategic Policy Forum on Digital entrepreneurship 2016, WEF 2018). In 2016 WEF claimed that in the next five years a widespread disruption in business models but also to labour markets, with enormous change predicted in the skill sets needed to thrive in the new landscape (WEF 2016). In an EU funded project these converging socio-technical phenomena have been studied and a framework has been proposed to analyse (and eventually design) work along the two complementary dimensions of workers’ competences and workplace properties. After having used this framework to explore the changes of work in some organizational contexts (eg industrial plants, network marketing), we are currently focusing on the competences of managers, since they are key in the design and deployment of digital transformation projects.

KEYWORDS: digital transformation, competences

References

Strategic Policy Forum on Digital Entrepreneurship, “Accelerating the digital transformation of European industry and enterprises”, March 2016.

Wang, J., & Verma, A. “Explaining organizational responsiveness to work-life balance issues: The role of business strategy and high-performance work systems”. Human Resource Management, 51(3), 407-432, 2012.

WEF (World Economic Forum). “The future of jobs: employment, skills and workforce strategy for the fourth industrial revolution.” World Economic Forum, Geneva, Switzerland, 2016.

ANALYSING INTERNATIONAL STUDENT MOBILITY FLOWS IN HIGHER EDUCATION. A COMPARATIVE STUDY ON EUROPEAN COUNTRIES

Marialuisa Restaino[1], Maria Prosperina Vitale[1] and Ilaria Primero[1]

[1] University of Salerno
(e-mail: mlrestaino@unisa.it, mvitale@unisa.it, iprimerano@unisa.it)

The study of international student mobility flows across different European countries has become an important research topic due to the relevance of internationalisation process in the university context. The analysis of the factors pulling and pushing students in a foreign country to complete higher education is, indeed, a key feature for the implementation of university policies in order to increase the number of ECTS gained abroad.

In line with related studies (Breznik et al., 2013; Barnett et al., 2016; Kondakci et al., 2018), the present contribution aims at identifying the characteristics of the student mobility trajectories involved in the Erasmus programme by considering a network analysis approach. Starting from this theoretical and analytical perspective, the main purposes are to discover the role played by each country revealing the presence of national hubs (i.e. good exporting countries) and authorities (i.e., good importing countries) and to explore the global network pattern identifying core-periphery or other topological structures in the destinations of Erasmus students. Thanks to the European Union Open Data Portal (EU ODP), a statistical overview of Erasmus mobility for students from 2008/09 to 2013/14 is obtained. The raw data are presented at country and European level including students' characteristics, among others age, gender, duration, subject area, level of study, sending and receiving country. From the EU ODP portal, at macro-level perspective, temporal network data structures (i.e. weighted and directed one-mode networks) are defined in which the nodes are the countries and the links represent the student mobility exchange between them with a weight proportional to the number of students involved. Hence, the directed networks are built considering the outgoing students and the incoming students.

KEYWORDS: Directed and weighted network, Erasmus student mobility, European open data, Social Network Analysis

References

BARNETT, G. A., LEE, M., JIANG, K., & PARK, H. W. 2016. The flow of international students from a macro perspective: a network analysis. *Compare: A Journal of Comparative and International Education*, **46**, 533-559.

BREZNIK, K., SKRBINJEK, V., LAW, K., & DAKOVIC, G. 2013. On the Erasmus student mobility for studies. In Management, Knowledge and Learning International Conference (pp. 13-21).

KONDAKCI, Y., BEDENLIER, S., & ZAWACKI-RICHTER, O. 2018. Social network analysis of international student mobility: uncovering the rise of regional hubs. *Higher Education*, **75**, 517-535.

A DAILY SENTIMENT INDEX ON IMMIGRATION OF ITALIAN-SPEAKING TWITTER USERS

Righi A.[1], Bianco D.M.[2] and Gentile M.M.[3]

[1] Department for statistical production, Italian National Institute of Statistics, (e-mail: righi@istat.it)

[2] (e-mail: domenicom.bianco@gmail.com)

[3] (e-mail: mauro.gentile@iese.net)

Tweets, Facebook or Instagram are relevant sources to study migration and mobility allowing to analyse the events, their evolution, and their perceptions due to their accessibility and their ability to catch dynamic reactions. Online social media data can be processed aimed at identifying in real-time and in an unsolicited way the opinion of people in host countries toward immigrants or related topics.

During the refugee crisis that has affected the Mediterranean area since 2015, the media have highlighted links between the growth of arrivals and the increase of emotional attitudes of public opinion and they have also contributed to the rise of this emotionality using a narrative that has not always been neutral and that in some countries has contributed to determining negative attitudes towards migrants (Coletto et al, 2016). The political debate in Europe has been strongly affected by the refugees crisis. The increase of immigration flows in 2015 and 2016 created both strong divergences among the EU countries on how to deal with the arrivals and arising problems in terms of social integration of newcomers. Social media hosted a harsh debate because the public opinion was divided between solidarity towards the refugees and refusal of the reception of migrants for economic reasons.

After having reviewed some previous research experiences on the topic, the data and the methods used for the construction of a daily sentiment index on migrations of Italian-speaking Twitter users is described and main results on the polarity of the sentiment and his evolution over the time are presented. The discussion on opportunities and limits of data and methods concludes the work.

KEYWORDS: 'social media', 'migration', 'sentiment analysis'.

References

COLETTO, M., LUCCHESE, C., MUNTEAN, C.I., NARDINI, F. M., ESULI, A., RENSO, C. AND PEREGO, R. (2016B). Sentiment-enhanced multidimensional analysis of online social networks: Perception of the Mediterranean refugees crisis, *ASONAM 2016*, San Francisco, California

EMOTIONOGRAPHIES OF THE CITY OF MILAN

Elisabetta Risi[1]

[1] Department of Communication, arts and media, IULM University of Milan
(e-mail: elisabetta.risi@iulm.it)

The paper aims to reconstruct a mapping of the emotional experiences in the city of Milan, and in some specifics places in particular, based on the analysis of online users generated contents.

The research project (still in progress) was implemented through the web crawling and the analysis of both textual and visual contents.

Today, social media are indispensable sources for monitoring opinions and sentiments expressed by citizens (Ceron et al., 2014). In the study, the contents of two social networking sites and one online review site were scraped and analysed (for a sort of research triangulation).

In particular, it was carried out:

- a sentiment analysis and opinion mining (Balahur & Jacquet, 2015; Younis, E. 2015) of the reviews of users who have been published on TripAdvisor, analysing the pages associated with specific places in the city
- object detection and recognition (Beier et al, 2012; Richards & Tuncer, 2018.) in images that are shared on Instagram, selecting only photo that have some tags (#milano #parcosempione #brera etc.)
- social media mining (Zafarani et al., 2014) of some selected Twitter accounts and tweets that have a specific tagging (for example those connected to Milan's public transport) scraped through Socioviz.

In the project are combined and compared different techniques, models and software of text mining (R and Python) and classification of images (Rapid Miner and Pattern) using types of supervised and unsupervised algorithms.

KEYWORDS: social media mining, sentiment analysis, image recognition

References

BAIER, D., DANIEL, I., FROST, S., NAUNDORF, R. 2012. Image data analysis and classification in marketing. *Advances in Data Analysis and Classification*, **6** (4), 253-276.

RICHARDS, D. R. & TUNÇER, B. 2018. Using image recognition to automate assessment of cultural ecosystem services from social media photographs, *Ecosystem Services,* **31**, 318-325.

BALAHUR, A., JACQUET G., 2015. Sentiment analysis meets social media – Challenges and solutions of the field in view of the current information sharing context, *Information Processing & Management,* **51**, 428-432.

ZAFARANI, R., ABBASI, A. M, & LIU H. 2014. *Social Media Mining: An Introduction*. Cambridge University Press.

YOUNIS, E. 2015. Sentiment Analysis and Text Mining for Social Media Microblogs using Open Source Tools: An Empirical Study. *International Journal of Computer Applications*, 112, 44 -48

HUMAN RESOURCES IN NEUROMANAGEMENT: AN ADDED VALUE TO JOB ASSESSMENT

Vincenzo Russo, Margherita Zito and Giorgio Gabrielli

Department of Business, Law, Economics and Consumer Behaviour "Carlo A. Ricciardi"
Università IULM,
(e-mail: Vincenzo.russo@iulm.it, margherita.zito@iulm.it, giorgio.gabrielli@iulm.it)

Human Resources have to find the right people for the job, and attract/identify talents. This requires the ability to communicate a good company brand image, and to understand candidates' characteristics and potential. While recruitment specialists have specific strategies for the latter, the first goal is often ignored.

As for methodological aspects, a neuroscientific approach measured the candidates' experience during a job interview. Thirty participants took part individually to a real job interview lasted forty minutes and emotional activation was measured in real time through skin conductance sensors (SC) (emotional arousal index), and an ElectroEncephaloGram (EEG) headset (engagement index) (Bolls et al. 2001).

Data showed that a) the most stressful parts of the interview such as agency mandate (SC mean=35.87; sd=37.54) and explanation of the work (SC mean=33.95; sd=35.57); b) the parts of the interview characterized by engagement such as career aspects (EEG mean=0.499; sd=0.65) and company presentation (EEG mean=0.485; sd=0.49).

Limitations of this study could be the use of a unique work context, but this study is a valid start point to detect also wellbeing and performance during assessment sessions. Moreover, even if these are data collected on a small sample, they have the potential of being shared and organized in longtail dataset, becoming a source of knowledge and vast applicability (Ferguson et al., 2014). Also, data based on small part of participant, can increase and improve study focusing on generic population, leading to different perspective in the study linked to specific issues (Wallis et al., 2013).

This research could help Human Resources to define "best practices" in terms of verbal and non-verbal language to put candidates at their ease, increasing their motivation to join the company. Moreover, this study is functional to suggest the interview strategies reducing the candidates' stress level and improving engagement. Future studies should provide the application of this scientific approach to advance employee training and to support career development.

KEYWORDS: 'Human Resources', 'Neuromanagement', 'Small samples'

References

BOLLS, P. D., LANG, A., & POTTER, R. F. (2001). The effect of message valence and listener arousal on attention, memory, and facial muscular responses to radio advertisements. *Communication Research*, **28**(5), 627–651.

FERGUSON, A. R. NIELSON, J. L., CRAGIN, M. H., BANDROWSKI, A. E., & MARTONE, M. E. (2014). Big data from small data: data-sharing in the 'long tail' of neuroscience. *Nat Neurosci*. **17**(11): 1442–1447.

WALLIS, J. C., ROLANDO, E. & BORGMAN, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PlosOne, **8**, 1-17.

“The Theatre Of The Mind”: The Effect Of Radio Exposition On Tv Advertising

Vincenzo Russo, Alessandro Fici, Margherita Zito and Giorgio Gabrielli

Department of Business, Law, Economics and Consumer Behaviour “Carlo A. Ricciardi”
Università IULM,
(e-mail: Vincenzo.russo@iulm.it, alessandro.fici1@gmail.com, margherita.zito@iulm.it, giorgio.gabrielli@iulm.it)

The radio has the power to create pictures in listeners’ mind, stimulating them to “see” what they are hearing (Bolls, 2002). The ability to evoke a sort of visual perception during the listening of an advertisement implies the possibility to engage the listeners as visual exposure to the product. This study aims to verify whether there is a difference in visual behavior on brand and in the emotional response between consumers exposed to a television advertisement or a website banner and consumers’ exposed to the same advertisement who first listened it via radio.

Seventy subjects participated in a between-subjects experiment in two conditions: commercial TV and web banners. In both conditions, half of the participants were first exposed to the same advertisements via radio. During the experiment, participants’ eye movements were recorded through an Eye-Tracker and their emotional activation was measured by skin conductance sensors (index of emotional arousal) and an EEG headset (index of Approach-Avoidance), recording alpha wave through the sensors in the F3 and F4 placement of the SI 10-20 (Cacioppo et al., 2000). Results showed that participants who were exposed to the radio advertisements before the watching of the television advertisements had a higher skin conductance level and a higher FAA index than the other participants. This pattern of results reflects a more intense and positive emotional response to the television advertisements if participants previously listened the same advertisements via radio. Moreover, results showed that the group of participants previously exposed to the advertisements via radio had a greater visual attention to the brand in the same advertisements watched on television. The same pattern of result was found for the web banner condition.

Within neuromarketing literature, this is not a very small sample, but it could be enlarged in order to have more generalizable data. Even if this, data based on restricted databases, can be shared and organized in longtail dataset. Heterogeneous data can be read in the light of a different source of knowledge (Ferguson et al., 2014), offering a different and integrating interpretation of reality in the study of a specific issues (Wallis et al., 2013).

Results highlight the possibility to improve advertising performances acting on other media, especially radio. Indeed, our research evidences that radio advertisements have a strong effect on web and television advertisements in terms of more visual attention on brand and higher consumers’ engagement, supporting the usefulness of a media-mix comprising the radio and suggesting that creating a multichannel

campaign with radio advertisements first can help to realize a more successful campaign.

Keywords: ‘Radio Power, ‘Neuromanagement’, ‘Small samples’

References

Bolls, P. D., Lang, A., & Potter, R. F. (2001). The effect of message valence and listener arousal on attention, memory, and facial muscular responses to radio advertisements. *Communication Research*, **28**(5), 627–651.

Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (2000). *Handbook of psychophysiology* (2nd ed.). New York: Cambridge University Press.

Ferguson, A. R. Nielson, J. L., Cragin, M. H., Bandrowski, A. E., & Martone, M. E. (2014). Big data from small data: data-sharing in the ‘long tail’ of neuroscience. *Nat Neurosci*. **17**(11): 1442–1447.

Wallis, J. C., Rolando, E. & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PlosOne, **8**, 1-17.

A COMPARISON OF MACHINE LEARNING TECHNIQUES FOR UNIVERSITY STUDENTS DROPOUT PREDICTION

Annalina Sarra[1], Eugenia Nissi[1] and Lara Fontanella[2]

[1] Department of Economics, University of "G.d'Annunzio" of Chieti-Pescara, (e-mail: asarra@unich.it, nissi@unich.it)
[2] Department of Legal and Social Sciences, University of "G.d'Annunzio" of Chieti-Pescara (e-mail: lfontan@unich.it)

Over the past two decades a large number of research studies attempt to predict the dropping out of students from university. Dropout phenomenon affects at every level the educational systems of any countries and represents a large concern to the education community and policy makers (Tinto 2006). Because of its consequences at social, institutional and personal level (Ulriksen et al. 2010), there is no doubt that dropout is a challenging problem. Student dropout is also a complex phenomenon because there is a broad array of factors that influence a student's likelihood of interrupting his/her educational courses. Thus, it becomes important to develop methods that facilitate prediction of students at risk of academic failure.

In this study we carried out a predictive analysis by relying on Educational Data Mining tools (Vandamme et al. 2007)

In order to classify students more likely to dropout we selected and compared some widely used classification techniques. In particular, Decision Trees, Artificial Neural Networks, Naïve Bayes classifiers and Bayesian Networks, have been chosen for their reliability, efficiency and popularity in the research community. We provide an illustration of the performance of these techniques through the employment of data from administrative repository, concerning students enrolled in the undergraduate courses of University of Chieti-Pescara.

KEYWORDS: academic dropout, educational data mining, machine learning techniques.

References

TINTO, V. 2006 Research and practice of student retention: what next? *Journal of College Student Retention: Research, Theory & Practice*, **8**(1):1–19.

ULRIKSEN, L. MADSEN, L.M. & HOLMEGAARD, H. T. 2010 What do we know about explanations for drop out/opt out among young people from STM higher education programmes? *Studies in Science Education,* **46**(2):209–244.

VANDAMME, J.P., MESKENS, N. & SUPERBY, J.F. Predicting academic performance by data mining methods. *Education Economics*, **15**(4):405–419.

PERSONALIZED TREATMENT OF CHRONIC DISEASES: MODELS, EXPERIMENTS AND PRELIMINARY RESULTS

Fabio Sartori[1], Riccardo Melen[1], Matteo Lombardi[1] and Davide Maggiotto[1]

[1] Department of Informatics, Systems and Communication, University of Milano-Bicocca,
(e-mail: fabio.sartori@unimib.it, riccardo.melen@unimib.it, matteo.lombardi@unimib.it, davide.maggiotto@unimib.it)

Quality of life (QoL) of patients affected by chronic diseases and their caregivers is a very important and interdisciplinary research topic. From recent literature (Adelman et al., 2014 new methodologies to reduce the impact of a chronic disorders on everyday life of affected people and their relatives are required: PERsonal Care Instructor and VALuator (PERCIVAL) project is an attempt to build up an integrated environment to promote the sharing, deliberation and monitoring of decisions about different aspects of chronic diseases among all the actors involved. The PERCIVAL project aims at developing an integrated, sharable and real-time model of knowledge and information involved in the treatment of chronic diseases. Patients affected by chronic diseases are often interested by multiple disorders, which make them frail from many points of view: From the physical perspective, they must follow different kinds of therapies, with possible negative intersections that could cause them very hard side-effects; From the psychological standpoint, the self-acceptation of such long-term disability, and, for most of them, life-long disability, is very difficult to reach; Last, but not least, from the social viewpoint, they must entirely depend on their caregivers, that are their relatives most of times, for all the aspects related to their condition; this dependency typically becomes closer and closer according to the evolution of the chronic disease. PERCIVAL aims at developing decision support systems capable to exploit both quantitative data perceived by wearable devices, that are nowadays recognized as very useful to face with chronic disorders (Rovini et al., 2017) and qualitative data provided by the user to take decisions tailored on patient profile. As a consequence, the PERCIVAL project aims at supporting virtual communities of practice involved in the chronic disorder management, allowing them to share data, documents and decisions. In this way, the overall QoL of more fragile subjects, in particular the caregivers, should increase thanks to the perception of being co-participants in chronic disease management, rather the main participants as usual. In this paper, we present the general architecture and a prototype of the PERCIVAL system focusing on the definition of personalized training programmes. Moreover, we describe the conceptual model behind the system, based on the adoption of Bayesian networks and the first experiments, whose goal was to profile potential users of the PEERCIVAL system to derive effective Conditional Probability Tables.

Keywords: wearable devices, bayesian networks, decision support systems.

References

Adelman, RD, Tmanova, L, Delgado, D., Dion, S. & Lachs, M. 2014. Caregiver burden: a clinical review. Jama 311, 10 (2014), 1052–1060.

Rovini,E., Maremmani, C. & Cavallo, F. 2017. How Wearable Sensors Can Support Parkinson's Disease Diagnosis and Treatment: A Systematic Review. Frontiers in neuroscience 11 (2017), 555.

SOCIAL MEDIA BIG DATA: STATE OF THE ART OF SOME METHODOLOGICAL CHALLENGES

Sciandra A.[1]

[1] STAR.Lab - Socio Territorial Analysis and Research, University of Padova, (e-mail: andrea.sciandra@unipd.it)

The increasing scientific production about social media Big Data brings with it several methodological issues outlined in literature. This paper aims at taking stock of the progresses made in social science research with Big Data, especially coming from social media, in the following processes: access, validity assessment and analysis. Focusing on data access, given the current restrictions on the APIs use, a possible solution is to get back to collect data with user consent (Bechmann et al., 2015) or to exploit new types of partnership between academia and private industry. Representativeness has often been a sore point in research using APIs to retrieve data from social media. However, some authors have shown that these limits can be overcome by improving the sample population coverage (Sampson et al., 2015). Other issues concerning noise detection in Big Data have been tackled with new methods, for instance in determining the presence of bots (Chu et al., 2010). Finally, in terms of Big Data analysis, supervised machine learning with human coding (Ceron et al., 2017) represents a fundamental technique to reduce the bias, partially recovering the link with social theory and stressing once again the need for an interdisciplinary approach.

KEYWORDS: social media, big data, access, validity, supervised machine learning.

References

BECHMANN, A., & VAHLSTRUP, P.B. 2015. Studying Facebook and Instagram data: The Digital Footprints software. *First Monday*., **20**(12).

CERON, A., CURINI, L., & IACUS, S.M. 2017. *Politics and Big Data: Nowcasting and Forecasting Elections with Social Media*. New York: Routledge.

CHU, Z., GIANVECCHIO, S., WANG, H., & JAJODIA, S. 2010. Who is tweeting on Twitter: human, bot, or cyborg?. In *Proceedings of the 26th annual computer security applications conference*., 21-30.

SAMPSON, J., MORSTATTER, F., MACIEJEWSKI, R., & LIU, H. 2015. Surpassing the limit: Keyword clustering to improve Twitter sample coverage. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*., 237-245.

CLASSIFICATION SUPERVISED MODELS TO DISCLOSE ACTION AND INFORMATION IN "U.N. 2030 AGENDA" SOCIAL MEDIA DATA

Sciandra A.[1], Finos L.[2] and Surian A.[3]

[1] STAR.Lab - Socio Territorial Analysis and Research, University of Padova, (e-mail: andrea.sciandra@unipd.it)

[2] Department of Developmental Psychology and Socialisation, University of Padova, (e-mail: livio.finos@unipd.it)

[3] Department of Philosophy, Sociology, Education and Applied Psychology (FISPPA), University of Padova, (e-mail: alessio.surian@unipd.it)

In 2015, the United Nation General Assembly adopted the 2030 Agenda for Sustainable Development (UN G.A., 2015) and its 17 Sustainable Development Goals (SDGs) aiming at ending all forms of poverty, fighting inequalities and tackling climate change. We collected data about the 2030 Agenda from May 9th to November 9th, 2018. The resulting dataset consists in a corpus of N = 209216 tweets. The aim of this work is to obtain a classification of each tweet in the corpus according to the "Information" - "Action" categories, in order to detect whether a tweet refers to an event or it has only an informative-disclosure purpose. Explicit intention to act or inform had been captured by hand coding of a randomly selected sample of tweets and then the classification had been extended to the whole corpus through different supervised machine learning methods. A textual analysis allowed us to include some variables related to SDGs hashtag and also the tf-idf weighting, to show how important a word is to distinguish "Information" and "Action" tweets in our corpus. Although we are aware of the limitations of this kind of studies (representativeness, noise, bots, etc.), we believe that probabilistic short-term models applied to the empirical observation of human behaviour in large datasets mined from social media could extract social knowledge, representing a great opportunity for social scientists (Lauro et al., 2017).

KEYWORDS: machine learning, textual data analysis, sustainable development goals

References

LAURO, N.C., AMATURO, E., GRASSIA, M.G., ARAGONA, B., & MARINO, M. (EDS.) 2017. *Data Science and Social Research: Epistemology, Methods, Technology and Applications*. Heidelberg: Springer.

UNITED NATIONS GENERAL ASSEMBLY 2015. *Transforming our world: the 2030 Agenda for Sustainable Development*, Resolution 70/1

PROTECT YOUR BUSINESS: A MODEL OF PROFILING SMALL BUSINESSES AND FREELANCERS ON ABILITY TO SECURE ONLINE

PROTEGGI LA TUA ATTIVITÀ: UN MODELLO DI PROFILAZIONE DELLE PICCOLE IMPRESE E DEI LIBERI PROFESSIONISTI PER ATTITUDINE AD ASSICURARSI ONLINE

Bruno Giuseppe Sfogliarini[1]

[1] Department of Marketing, Communication and Consumer Behavior, University IULM, via Carlo Bo, Milan, Italy
e-mail: bruno.sfogliarini@iulm.it

This paper provides a solution by text mining techniques to profile small businesses and freelancers respect to their attitude to subscribe an insurance coverage, to protect them from risks relate to their professional activity, provided by an online broker. The raw data adopted to estimate the profiling model are drawn from public personal profiles as published on Linkedin social network, complemented by an original primary research conducted on a representative sample of Italian enterprises and professionals, covering a comprehensive selection of economic activities.

KEYWORDS: Professional insurance, profiling method, text mining, small business, freelance, Linkedin, market research

MATCHING ON POSET BASED AVERAGE RANK FOR MULTIPLE TREATMENTS (MARMOT)

Silan, M. [1], Boccuzzo, G.[1] and Arpino, B.[2]

[1] Department of Statistical Sciences, University of Padua, Padua, Italy, (e-mail: silan@stat.unipd.it, boccuzzo@stat.unipd.it)

[2] Universitat Pompeu Fabra, Barcelona, Spain (e-mail: bruno.arpino@upf.edu)

In a multi-treatment framework, the use of propensity score techniques needs a different specification and a peculiar attention. The generalised propensity score (GPS) is the conditional probability of receiving a particular level of the treatment given the pre-treatment variables in a multiple-treatment framework. However, applications of generalised propensity score remain mainly scattered in literature, with few applications in three (or four) treatments regimes, and existing methods are highly unpractical and nearly impossible to perform if the number of treatments increases greatly. Indeed, some important assumptions, such as the overlap, are difficult to satisfy and the estimation of the propensity score becomes computationally heavy.

We propose an original alternative approach to deal with covariate balance in the context of the estimation of the causal effect of many treatments. Our method consists of matching on a score (average rank) obtained using the theory of Partially Ordered Sets (poset). This approach that we label MARMoT (Matching on Poset based Average Rank for Multiple Treatments) allows making the distribution of confounders similar across many treatments.

The aim of this work is to estimate the neighbourhood effect in the city of Turin using a propensity score matching approach to adjust for confounders. Before doing so, we tested with a simulation study the MARMoT approach. The main idea that underlies this technique is to obtain a population in which each profile, each combination of confounders summarized by the poset based average rank, is equally represented in all the treatment groups. MARMoT technique has proved to be really useful to balance for counfounders and reduce biases in estimates. Moreover, the matching is not bounded to subjective choices and the computational time is limited even in the case with 70 different treatments.

KEYWORDS: ”Matching”, ”Multi-treatment”, ”Neighbourhood effect”, ”Poset”.

BLENDING SMALL AND BIG DATA FOR EVALUATING THE IMMIGRATION-INSECURITY RELATIONSHIP

Sonia Stefanizzi[1] and Giancarlo Manzi[2]

[1] Department of Sociology and Social Research, University of Milan-Bicocca, Italy,
(e-mail: giancarlo.manzi@unimi.it, silvia.salini@unimi.it)

[2] Department of Economics, Management and Quantitative Methods, University of Milan, Italy,
(e-mail: giancarlo.manzi@unimi.it)

In recent years, perceived threat from immigration has been increasingly amplified by stereotyped visions obsessively pursued by populist media and parties through “easy arguments”: more immigration implies more crime, the national cultural life is undermined by immigrants, immigrants steal jobs to local population, etc. (Abrams et al. 2018; Bathia, 2018). This “aversion amplification” has fed important political shift in many countries in the EU such as the Brexit referendum in the UK and landslide wins of right-wing populist parties in national parliamentary elections in Hungary, Austria, Germany or Italy. In opinion survey, answers given to objective questions are often in contradiction with respect to answers to perception/opinion questions relating the same topic. Answers to sensitive questions regarding, say, racist attitudes, tend to mask the underlying real sentiment (Manzi & Saibene, 2018) due to moral reasons. Regarding the aforementioned “easy arguments”, the equation “more immigrants equal more crime” is often disavowed by data. In this paper we attempt to blend estimates of this relationship from “small” data sources with estimates obtained from big data sources. With “small data source” we mean data from traditional surveys, official or not. With “big data sources” we mean text data coming from social networks. We try to evaluate whether the difference between the estimates from these two data categories is significant or not and how it has evolved in the last years. Estimates from big data sources are obtained through sentiment analysis techniques. The blending exercise is performed by directly modelling the bias within the original data sources and using a hierarchical approach in a Bayesian framework.

KEYWORDS: Perceived threat of immigration, crime, Bayesian analysis, bias, hierarchical models.

References

ABRAMS, D. & TRAVAGLINO, G.A. 2018. Immigration, Political Trust, and Brexit – Testing An Aversion Amplification Hypothesis. *British Journal of Social Psychology*, **57**, 310-326.

BATHIA, M. 2018. Social Death: The (White) Racial Framing of the Calais 'Jungle' and 'Illegal' Migrants in the British Tabloids and Right-Wing Press. In Bathia, M., Poynting, S. and Tufail, W. (Eds.): *Media, Crime and Racism*, Springer.

MANZI, G., SAIBENE, G. 2018. Are they telling the truth? Revealing hidden traits of satisfaction with a public bike-sharing service. *International Journal of Sustainable Transportation*, **12**(4), 253-270.

THE REUSE OF DIGITAL COMPUTER DATA: TRANSFORMATION, RECOMBINATION AND GENERATION OF *DATA MIXES* IN BIG DATA SCIENCE

Niccolò Tempini[1]

[1] Egenis Centre for the Study of the Life Sciences, Department of Sociology, Philosophy and Anthropology, University of Exeter, and The Alan Turing Institute, London (e-mail: n.tempini@exeter.ac.uk)

This research is concerned with the relationship between the materiality of digital computer data and their reuse in scientific practice. It builds on the case study of a 'data mash-up' infrastructure developed at the crossroads between environmental and weather sciences and population health research. I illustrate the extent to which scientists reusing digital computer data heavily manipulate the sources through complex and situated calculative operation, as they attempt to re-situate data well beyond the epistemic community in which they originated, and adapt them to different theoretical frameworks, methods and evidential standards. The research interrogates the consequent relationship between *derivative* data and the data sources from which they originate. I argue that deep transformation and recombination of data source values and data structures, involved in the reuse of computer data, lead to the systematic creation of derivative data that are best considered new digital objects. In certain situations of scientific inquiry, data can be productively reused only if they are put in some kind of relation with other data, a relation which is realised in the computer relations stabilised by the construction of a new dataset object. The intense relationality of *scientific computer data* is multi-layered and scaffolded, as it depends on relations between various kinds of data, computing technologies, assumptions, theoretical scaffoldings, hypotheses and other features of the situation at hand. Understanding this, I argue, is crucially important to advance our understanding of the data reuse journeys of computer data.

KEYWORDS: computer data, scientific data, derivative data, data mash-up, data infrastructure.

References

HUI, Y. 2017. *On the Existence of Digital Objects.* Minneapolis, MN: University of Minnesota Press.
MANOVICH, L. 2001. *The Language of New Media*. Cambridge, MA: MIT Press.
SIMONDON, G. 2017. *On the Mode of Existence of Technical Objects.* Minneapolis, MN: University of Minnesota Press.

A PREFERENCE INDEX DESIGN FOR BIG DATA

Tomaselli Venera[1] and Cantone Giulio Giacomo[1]

[1] Department of Political and Social Sciences, University of Catania,
(e-mail: tomavene@unict.it, prgcan@gmail.com)

TripAdvisor is a business service that works as reputation system to guarantee quality in tourism experience. This kind of new services is based on Big Data technologies and characterized by generating, managing and summarizing, even with rating indexes, a quantitative experimental size of information, representing a frontier issue for data analysis.

These data are organized and offered to users by a filter system aimed to recommend consumer's choices. Through a methodological design oriented to reward competitive quality, this acts as a crowd-sourced evaluation system.

The stakeholders don't take the crowd-sourced evaluation passively as they still have a wide range of, reactive actions even with huge amounts of received reviews. They can re-organize their economic game of exposure to crowd-sourced reputation by adopting merely strategic and unsubstantial behaviours, or even frauds.

As past reviews and rating indexes provided by the websites can affect the building data process, in this paper we suppose that final information can be biased, leading itself into non-linear, asymmetric dynamics.

On the basis of an empirical study for approximately 26.000 scores on TripAdvisor multipoint scale organized into 8-years time series and harvested by *R* software, we propose a methodological design of a rating index. The index is robust for avoiding any manipulation of open-to-view-results multipoint scales and, at the same time, reflects the original aim to provide both a form of guarantee from risks of bad experience in tourism and a coherent ranking for benchmarking purposes.

KEYWORDS: crowd-sourced evaluation, TripAdvisor, multipoint scale, ranking system robustness.

References

MASUM, H., & TOVEY, M. 2011. *The Reputation Society: How Online Opinions Are Reshaping the Offline World*, Cambridge: MIT Press.

SAURO, J., & LEWIS, R. 2016. *Quantifying the User Experience. Practical Statistics for User Research, 2nd Edition*, Burlington: Morgan Kaufman, 249-276.

Measuring Immigrants Integration: Simultaneous Clustering and Dimensionality Reduction in PLS-SEM Models

Venera Tomaselli[1], *Mario Fardellone*[2], *Maurizio Vichi*[2]

[1] Department of Political and Social Sciences, University of Catania
[2] Department of Statistical Sciences, University "La Sapienza" Roma
(e-mail: tomavene@unict.it, mario.fordellone@uniroma1.it,
Maurizio.Vichi@uniroma1.it)

In the present study, due to the presence of nominal and ordinal variables, the models will be estimated by the Non-Metric PLSPM algorithm and, specifically, by a simultaneous non-hierarchical clustering and Partial Least Squares Modelling, named Partial Least Squares K-Means (PLS-KM (Fordellone and Vichi, 2017). 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 $\mathbf{\Lambda} = [\mathbf{\Lambda}_{\mathrm{H}}, \mathbf{\Lambda}_{\mathrm{L}}]$, and the errors matrices $\mathbf{Z}$ and $\mathbf{E}$, the Structural Equation Modeling *K*-Means approach can be written as follows:

$$\mathbf{X} = \mathbf{UC\Lambda\Lambda}^{\mathrm{T}} = \mathbf{UC\Lambda}_{\mathrm{H}}\mathbf{\Lambda}_{\mathrm{H}}^{\mathrm{T}} + \mathbf{UC\Lambda}_{\mathrm{L}}\mathbf{\Lambda}_{\mathrm{L}}^{\mathrm{T}} + \mathbf{E} \qquad (1)$$

under constraints: *(i)* $\mathbf{\Lambda}^{\mathrm{T}}\mathbf{\Lambda} = \mathbf{I}$; and *(ii)* $\mathbf{U} \in \{0,1\}$, $\mathbf{U1}_K = \mathbf{1}_n$. Thus, the SEM-KM model includes the SEM estimated via Partial Least Squares 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 through the optimization of the overall objective function.

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 Modeling (PLSPM) approach (Tenenhaus *et al.*, 2005).

Keywords: PLS-SEM, simultaneous clustering, immigrants integration index.

References

FORDELLONE, M., VICHI, M. (2018). *Structural Equation Modeling and simultaneous clustering through the Partial Least Squares algorithm*. *arXiv preprint arXiv:1810.07677*.

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

DESTINATION BRANDING AND RESIDENTS: AN ANALYSIS OF THE VALTELLINA TERRITORY

Giovanni Tonini[1], Mariangela Zenga[1], and Nadia Cominetti[1]

[1] University of Milano-Bicocca, Italy
(e-mail: giovanni.tonini@unimib.it; mariangela.zenga@unimib.it; n.cominetti@campus.unimib.it)

The residents' communication on their territory as tourist destination is becoming an important key in the place branding process. This paper aims to understand the relationship between the communication of the Destination Images and the word-of-mouth in the Valtellina territory (North Italy).

KEYWORDS: Destination Branding; Residents; Valtellina.

References

J. H. G. JEURING, T. HAARTSEN (2017) Destination Branding by Residents: The Role of Perceived Responsibility in Positive and Negative Word-of-Mouth, Tourism Planning & Development, 14:2, 240-259, DOI: 10.1080/21568316.2016.1214171

GEOCACHING – BOXING GEOSOCIAL TRACES

Judit Varga[1]

[1] School of Sociology and Social Policy, University of Nottingham, United Kingdom
(e-mail: judit.varga@nottingham.ac.uk)

This paper explores the normativity of social (big)data using multidisciplinary narratives – drawing on semi-structured interviews with researchers who use geotagged social media (geosocial) traces for scientific research - that depict instances of 'black-boxing' and 'infrastructural inversion' (Bowker & Star, 2000). Geosocial traces - such as geolocated tweets and Instagram posts – are used by scholars from diverse disciplines, such as computer science, geography, physics and sociology, to study situated practices – for example, people's activities in the neighbourhoods of a city. Through discussing how geosocial traces are positioned *as data* (cf. Venturini *et al*., 2018) to enable varied 'geographic imaginations', this paper contributes to two topics in STS. First, it helps STS to explore interdisciplinary data practices, beyond disciplinary enclaves (Edwards *et al*., 2011). Second, studying how *similar traces* get positioned as data *through various disciplinary – epistemic - cultures* can help understand diverse *values, valuations,* and *evaluations* of black boxing and infrastructural inversion, and the role of human and non-human aspects of assemblages thereof (Aragona, *et al.* 2018). Black boxing can enable scholars from diverse backgrounds to perform 'computational social science', which may have aesthetic, political and epistemic value for them. Infrastructural inversion can happen through friction – for example, when 'standard' data models don't 'fit' the data. On the other hand, some aspects of data assemblages, such as flickr users' skill or the 'everydayness' of Instagram posting are exposed and positioned as values that show the 'relevance' of research endeavours. Hence, infrastructural inversion isn't necessarily related to friction.

KEYWORDS: 'geotagged social media traces', 'big data', 'infrastructural inversion'

References

ARAGONA, B., FELACO, C., & MARINO, M. (2018). The Politics of Big Data Assemblages. *PArtecipazione e COnflitto,* **11**(2), 448–471.

EDWARDS, P. N., MAYERNIK, M. S., BATCHELLER, A. L., BOWKER, G. C., & BORGMAN, C. L. (2011). Science friction: Data, metadata, and collaboration. *Social Studies of Science,* **41**(5), 667– 690.

VENTURINI, T., BOUNEGRU, L., GRAY, J., & ROGERS, R. (2018). A reality check(list) for digital methods. *New Media & Society,* 1–23.

PART OF SPEECH TAGGING FOR BLOG AND MICROBLOGS DATA

Paola Zola [1] and Silvia Golia[1]

[1] Department of Economics and Management, University of Brescia, Italy
(e-mail: paola.zola@unibs.it, silvia.golia@unibs.it)

Part-of-speech (POS) tagging is at the basis of many Natural Language processing tasks. In the last decades, a huge number of researches has been done in this field and algorithms can reach high level of correct POS tags in a sequence. However, the increasing use of Internet and the explosion of blogs and microblogs changed the way people communicate and POS taggers, trained on structured corpora, lose ability to catch this new tendency (Nand and Perera, 2015). In fact, the most famous algorithms to POS tag are based on Markov Models and thus, they estimate the probability of an unknown tag given the knowledge of previous POS tags and the word's structure. The aim of this work is to predict the unknown POS tag only evaluating the sequence of POS tags, thus, avoiding the probability of a POS tag given the word itself. The statistical model used to reach this goal is the Bayesian Network (BN). The graphical structure of the BN is derived from the data though a score based algorithm identified thanks to a 10-fold cross validation analysis (Scutari and Denis, 2014). Then, the estimated BN is used to predict the probability distribution of the unknown POS tag and the predicted POS tag is the one corresponding to the mode of its distribution. The dataset over which this approach is applied is the well-known Brown Corpus, which is composed by more than one million token related to a huge range of topics as linguistics, psychology, statistics and sociology. Preliminary results have shown a sufficient ability of the BN to predict unknown POS tags. Moreover, the Random Forest is considered as an alternative approach to POS tag prediction; preliminary analysis on the Brown Corpus have shown poorly performance compared to BNs.

KEYWORDS: Microblogs, Bayesian Networks, Random Forest

References

NAND P & PERERA R 2015. An evaluation of pos tagging for tweets using hmm modeling.

SCUTARI M. & Denis J.B. 2014. Bayesian Networks with Examples in R.*Chapman and Hall/CRC*

The ”Sentability”: sentiment analysis for social media’s accountability. The case of environmental issues in Italian Municipailities.

Paola Zola [1], Laura Rocca[2],Davide Giacomini[1] and Diego Paredi[1]

[1] Department of Economics and Management, University of Brescia, Brescia, Italy (e-mail: paola.zola@unibs.it, davide.giacomini@unibs.it, d.paredi92@gmail.com)

[2] Department of Law, University of Brescia, Brescia, Italy (e-mail: laura.rocca@unibs.it)

Due to the expansion of the Internet and Web 2.0 phenomenon, there is a growing amount of freely opinionated text. Modern techniques in natural language processing and machine learning might be useful to investigate important aspects of the communication between different authors. In this paper we aim to analyze the sentiment of citizens related to environmental sustainability messages published by Italian Local Government on Facebook.An initial study, on our sample of 39 municipalities, is performed to detect possible subtopic applying Latent Dirichlet Allocation. Then, having a list of posts related to specific environmental sustainable theme we computed the sentiment of citizens comments (Ortigiosa et al., 2014).In this analysis we decided to focus on environmental topics because it is a key concept for global and local growth. Moreover, according to the Agenda 21 action, Local governments play a central role since their level of government is the closest to citizens daily lives (Gesuele, 2016).The main results show an increasing impact of Web2.0 in Local Governments direct interaction with citizens but, a divergence of interest in environmental topics between the two actors.

Keywords: Facebook, municipalities, NLP, text mining

References

GESUELE, B. 2016. Municipalities and Facebook Use: Which Key Drivers? Empirical Evidence from Italian Municipalities. *International Journal Of Public Administration*, **39(10)**, 771-777.

ORTIGIOSA, A., MARTÍN, J. M., and Carro, R. M. 2014. Sentiment analysis in facebook and its application to e-learning. *Computers in Human Behavior*, **31**, 527541.

Sponsors

University of Milano - Bicocca
Department of Economics, Management and Statistics
Department of Sociology and Social Research
Department of Informatics, Systems and Communication

Partners

Applied Statistics Association
Associazione Italiana di Sociologia
Associazione Italiana di Ricerca Operativa
Bicocca Applied Statistics Center
Consorzio Milano Ricerche
Italian Statistical Society
PKE
Unidata Bicocca

Data Science & Social Research 2019

Book of Abstracts

Editor: Paolo Mariani

The book includes the abstracts accepted for presentation at the Data Science & Social Research 2019 conference, February 4-5, in Milan. The conference aims at stimulating the debate between scholars of different disciplines about the so called "data revolution" in social research. Statisticians, computer scientists and domain experts in social research will discuss the opportunities and challenges of the social data revolution to create a fertile ground for addressing new research problems.

ISBN: 9788894312096