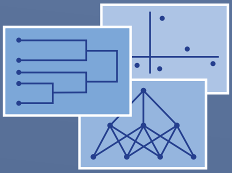
Studies in Classification, Data Analysis, and Knowledge Organization

N. Carlo Lauro · Enrica Amaturo
Maria Gabriella Grassia
Biagio Aragona · Marina Marino
Editors

Data Science and Social Research

Epistemology, Methods, Technology and Applications



Springer

Studies in Classification, Data Analysis, and Knowledge Organization

More information about this series at http://www.springer.com/series/1564

N. Carlo Lauro · Enrica Amaturo
Maria Gabriella Grassia · Biagio Aragona
Marina Marino
Editors

Data Science and Social Research

Epistemology, Methods, Technology and Applications

Springer

Editors
N. Carlo Lauro
Department of Economy and Statistics
University of Naples Federico II
Naples
Italy

Enrica Amaturo
Department of Social Sciences
University of Naples Federico II
Naples
Italy

Maria Gabriella Grassia
Department of Social Sciences
University of Naples Federico II
Naples
Italy

Biagio Aragona
Department of Social Sciences
University of Naples Federico II
Naples
Italy

Marina Marino
Department of Social Sciences
University of Naples Federico II
Naples
Italy

ISSN 1431-8814 ISSN 2198-3321 (electronic)
Studies in Classification, Data Analysis, and Knowledge Organization
ISBN 978-3-319-55476-1 ISBN 978-3-319-55477-8 (eBook)
DOI 10.1007/978-3-319-55477-8

Library of Congress Control Number: 2017937698

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG
part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Data Science is a multidisciplinary approach based mainly on the methods of statistics and computer science suitably supplemented by the knowledge of the different domains to meet the new challenges posed by the actual information society. Aim of Data Science is to develop appropriate methodologies for purposes of knowledge, forecasting, and decision-making in the face of an increasingly complex reality often characterized by large amounts of data (big data) of various types (numeric, ordinal, nominal, symbolic data, texts, images, data streams, multi-way data, networks, etc.), coming from disparate sources.

The main novelty in the Data Science is played by the role of the KNOWLEDGE. Its encoding in the form of logical rules or hierarchies, graphs, metadata, and ontologies, will represent a new and more effective perspective to data analysis and interpretation of results if properly integrated in the methods of Data Science. It is in this sense that the Data Science can be understood as a discipline whose methods, result of the intersection between statistics, computer science, and a knowledge domain, have as their purpose to give meaning to the data. Thus, from this point of view, it would be preferable to speak about DATA SCIENCES.

The Data Science and Social Research Conference has represented an interdisciplinary event, where scientists of different areas, focusing on social sciences, had the opportunity to meet and discuss about the epistemological, methodological, and computational developments brought about by the availability of new data (big data, big corpora, open data, linked data, etc.). Such a new environment offers to social research great opportunities to enhance knowledge on some key research areas (i.e. development, social inequalities, public health, governance, marketing, communication).

Along, the conference has been a crucial issue to discuss critical questions about what all this data means, who gets access to what data, and how data are analysed and to what extent.

Therefore, aim of the conference, and of the present volume, has been to depict the challenges and the opportunities that the “data revolution” poses to Social Research in the framework of Data Science, this in view of building a SOCIAL DATA SCIENCE … Let us own data science!

Naples, Italy

N. Carlo Lauro
Professor Emeritus of Statistics

Contents

Introduction 1
Enrica Amaturo and Biagio Aragona

Part I Epistemology

On Data, Big Data and Social Research. Is It a Real Revolution? 9
Federico Neresini

New Data Science: The Sociological Point of View 17
Biagio Aragona

Data Revolutions in Sociology 25
Barbara Saracino

Blurry Boundaries: Internet, Big-New Data, and Mixed-Method Approach 35
Enrica Amaturo and Gabriella Punziano

Social Media and the Challenge of Big Data/Deep Data Approach 57
Giovanni Boccia Artieri

Governing by Data: Some Considerations on the Role of Learning Analytics in Education 67
Rosanna De Rosa

Part II Methods, Software and Data Architectures

Multiple Correspondence K-Means: Simultaneous Versus Sequential Approach for Dimension Reduction and Clustering 81
Mario Fordellone and Maurizio Vichi

TaLTaC 3.0. A Multi-level Web Platform for Textual Big Data in the Social Sciences 97
Sergio Bolasco and Giovanni De Gasperis

Sparsity Data Reduction in Textual Network Analysis: An Exercise on Sustainability Meaning 105
Emma Zavarrone, Filomena Grassia, Maria Gabriella Grassia and Marina Marino

University of Bari's Website Evaluation 121
Laura Antonucci, Marina Basile, Corrado Crocetta, Viviana D'Addosio, Francesco D. d'Ovidio and Domenico Viola

Advantages of Administrative Data: Three Analyses of Students' Careers in Higher Education 131
Andrea Amico, Giampiero D'Alessandro and Alessandra Decataldo

Growth Curve Models to Detect Walking Impairment: The Case of InCHIANTI Study 141
Catia Monicolini and Carla Rampichini

Recurrence Analysis: Method and Applications 151
Maria Carmela Catone, Paolo Diana and Marisa Faggini

Part III On-Line Data Applications

Big Data and Network Analysis: A Promising Integration for Decision-Making 165
Giovanni Giuffrida, Simona Gozzo, Francesco Mazzeo Rinaldi and Venera Tomaselli

White House Under Attack: Introducing Distributional Semantic Models for the Analysis of US Crisis Communication Strategies 175
Fabrizio Esposito, Estella Esposito and Pierpaolo Basile

#Theterrormood: Studying the World Mood After the Terror Attacks on Paris and Bruxelles 185
Rosanna Cataldo, Roberto Galasso, Maria Gabriella Grassia and Marino Marina

Learning Analytics in MOOCs: EMMA Case 193
Maka Eradze and Kairit Tammets

Tweet-Tales: Moods of Socio-Economic Crisis? 205
Grazia Biorci, Antonella Emina, Michelangelo Puliga, Lisa Sella and Gianna Vivaldo

The Sentiment of the Infosphere: A Sentiment Analysis Approach for the Big Conversation on the Net 215
Antonio Ruoto, Vito Santarcangelo, Davide Liga, Giuseppe Oddo, Massimiliano Giacalone and Eugenio Iorio

The Promises of Sociological Degrees: A Lexical Correspondence Analysis of Masters Syllabi 223
D. Borrelli, R. Serpieri, D. Taglietti and D. Trezza

Part IV Off-Line Data Applications

Exploring Barriers in the Sustainable Microgeneration: Preliminary Insights Thought the PLS-PM Approach 239
Ivano Scotti and Dario Minervini

Individual Disadvantage and Training Policies: The Construction of “Model-Based” Composite Indicators 249
Rosanna Cataldo, Maria Gabriella Grassia, Natale Carlo Lauro,
Elena Ragazzi and Lisa Sella

Measuring the Intangibles: Testing the Human Capital Theory Against the OECD Programme for the International Assessment of Adult Competencies 261
Federica Cornali

Analysis of the Employment Transitions and Analysis of the Unemployment Risk in the Social Security Account Statements of the Patronato ACLI 269
D. Catania, A. Serini and G. Zucca

Integrated Education Microdata to Support Statistics Production 283
Maria Carla Runci, Grazia Di Bella and Francesca Cuppone

Latent Growth and Statistical Literacy 291
Emma Zavarrone

Erratum to: Recurrence Analysis: Method and Applications E1
Maria Carmela Catone, Paolo Diana and Marisa Faggini

Introduction

Enrica Amaturo and Biagio Aragona

One of the fundamental features of modern societies is the never-ending quantity of data they produce as direct and indirect effects of business and administrative activities, as well as the result of volunteer accumulation of information on the Internet by individuals who use the Web for social relationships and knowledge construction. Changes in social networking and the pervasive use of the Web in daily life, as well as improvements in computational power and data storage, are having impressive effects on data production and consumption. Social networks, sensors, and data infrastructure are generating a massive amount of new data (big data, big corpora, linked data, open data, etc.) that are readily available for the analysis of societies. That is why some talked about a "data deluge" (The Economist 2010) able to radically change both the individual and the social behaviours, and others (Kitchin 2014) have labelled the present time as "the data revolution era". Data revolution is the sum of disruptive social and technological changes that are transforming the routines of construction, management, and analysis of data once consolidated within the different scientific disciplines.

Digital technology for scanning, processing, storing, and releasing data has already had an impact on the quality of information available to social researchers. Other opportunities are opened by computational changes that have a radical effect on the nature of dissemination by allowing to deal with large data even for small areas (Uprichard et al. 2008) and to make data storage possible also to individuals and small businesses. The wide availability of software for analysis also has to be considered when drawing the picture of the new possibilities offered by technological changes that affect the production and consumption of data (Baffour et al. 2013).

E. Amaturo · B. Aragona (✉)
Department of Social Sciences, University of Naples Federico II,
Vico Monte di Pietà, 1, Naples, Italy
e-mail: aragona@unina.it

E. Amaturo
e-mail: amaturo@unina.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_1

A first epistemological consequence of the passage from data scarce to data-intensive societies has been the re-emergence of data-driven science, which is opposed to hypotheses-driven science that is typical of post-positivist social science. The main argument of those who proclaim the "data first" model of science is that, being able to track human behaviour with unprecedented fidelity and precision, exploring existing data may be more useful than building models of why people behave the way they do. More specifically, in 2009, Lazer—in a paper that had great success within the scientific community (more than 1.527 citations)—identified Big Data as the core of a new field of social science which makes intensive use of computer science (computational social science (CSS)). For him, these vast data sets on how people interact were offering new perspectives on collective human behaviour.

As the availability of big quantities of data has grown, the main traditional empirical basis of quantitative social sciences (surveys and experiments) is being dismantled in favour of new data analysis. Market research, for example, widely employs studies on network communities instead of traditional survey's campaign and network, and sentiment analysis is substituting for traditional election pools, which proved to be less effective than they were in the past. Not to mention how documents' analysis has really changed with the advent of the Web and of social media (Amaturo and Aragona 2016). Because nearly all of our activities from birth until death leave digital traces in large databases, social scientist, who had to rely on account of actions for their research (through questionnaire), using new data can be in the action without asking questions or being seen.

After the early enthusiasm about the data deluge, in the past years critical data studies have been carried out to more deeply understand what is the context of validity of new data. Special attention has been paid, for example, to voluntarily generated contents on social network and Websites. They represent a massive quantity of data, but they need to be contextualized; otherwise, it becomes difficult to make sense of them. Moreover, despite the often made claim that Big Data provides total populations, ending our reliance on samples, this is rarely the case for social media data (Highfield et al. 2013). Boyd and Crawford, for example, have noted that working with Twitter data has: "serious methodological challenges that are rarely addressed by those who embrace it" (2012: 13) and that "Twitter does not represent people and it is an error to assume people and Twitter users as synonymous: they are a very particular sub-set" (2012: 12). When using data coming from the Web, researchers must recognize that part of the population is not accessible because does not have access to the Internet and that many are passive consumers of Internet information rather than active participants in the Web 2.0. Access may also be segmented according to socio-demographic characteristics (nationality, age, gender, education, income, etc.), systematically excluding some strata of the population from research. Surveys in the USA, for instance, show that Twitter has a disproportionate number of young, male black, and Hispanic users compared to the national population (Duggat et al. 2015).

These methodological concerns about validity and coverage biases rise also more deep sociological and political questions about to what extent these data may be used for the analysis of society, what aspects of the social reality they capture, and how they can be customized for designing, implementing, monitoring, and evaluating social policies. *La gouvernance par les nombres* (Supiot 2016) may be crucial to understand what will be the future of new data within both social sciences and society. Indeed, new digital data have been "normalized" within administrations, as showed by proliferating database-related technologies of governance. They are complementing existing uses of data with methods of digital governance, whereby digital technologies, software packages and their underlying standards, code, and algorithmic procedures are increasingly being inserted into the administrative infrastructure of our societies.

One interesting example is the fact that administrations, through local statistical offices, are giving access to their micro-data and have started to finance open-data initiatives with both a cognitive and a normative intent. On one side, open data help technicians, administrators, and politicians to redirect policies and, on the other side, allow citizens to check whether policies have had the desired impact. Opening data is therefore a consequence of the importance of transparency and accountability in our societies.

Another example concerns how Big Data have captured the interest of National Statistical Institutes (NSI) and related agencies such as Eurostat and the European Statistical System (ESS), who have formulated a Big Data roadmap. United Nations Economic Commission for Europe (UNECE) has established a High Level Group for the Modernization of Statistical Production and Services focused on Big Data with four "task teams": privacy, partnerships, sandbox and quality. Even the United Nations Statistical Division (UNSD) has organized a Global Working Group on Big Data and Official Statistics. The interest of official statistics is due to the fact that the developments in ICT help to handle these data sources and hence allow to drastically reduce the costs of statistics. However, a survey jointly conducted by UNSD and UNECE revealed that of the 32 NSIs that responded only a "few countries have developed a long-term vision for the use of Big Data", or "established internal laboratories, task teams, or working groups to carry out pilot projects to determine whether and how Big Data could be used as a source of Official Statistics" (EUESC 2015: 16).

A part from the efforts that NSI are doing in inserting Big Data in their statistical production, the use of Big Data, both structured and unstructured, can represent a valuable way to inspire decision-making at all level of public administration in a time of scarce resources. The technological revolution is in fact enabling governments to use a great variety of digital tools and data to manage all phases of the policy cycle's process more effectively, becoming a core element for e-governance applications and techniques. It has been widely claimed that this radical expansion of digital data is transforming the global evidence base and will lead to improved knowledge, understanding, and decision-making across the economy, in turn improving life chances and well-being for individuals and for the health and sustainability of economies and societies more broadly (Mayer-Schönberger and Cukier 2013; Margetts and Sutcliffe 2013). However, still more research is needed about what kind of analytics can be usefully managed, at what policy level they are really demanded, how they are collected, organized, integrated, and interrogated, by whom and for what purposes. Data are not useful in and of themselves. It is what is done with data that is important, and making sense of new data poses new analytical challenges.

First of all, new data usually require more attention to the processes where data have to be pre-prepared for analysis through data selection, curation, and reduction activities. Pre-analytical work can be extremely hard and time-consuming, so data scientists are devoting more research to seek the most productive, efficient and effective ways to undertake and especially, to automate, this work. Furthermore, the analysis of very large numbers of data records can be timely run only by computer algorithms, and then, much work is about developing automated processes that can assess and learn from the data and their analysis, the so-called machine learning. Machine learning seeks to iteratively evolve an understanding of datasets and is been used for data mining in order to detect, classify, and segment meaningful relationships, associations, and trends between variables. Data mining may employ a series of different techniques including natural language processing, neural networks, decision trees, and statistical (nonparametric as well as parametric) methods. The selection of techniques varies according to the type of data (structured, unstructured or semi-structured) and the objective of the analysis. Unstructured data in the form of natural languages raise particular data mining challenges; they need semantics and taxonomies to recognize patterns and extract information from documents. A typical application of such technique is sentiment analysis which seeks to determine the general nature and strength of opinions about an issue.

Another analytical challenge is about data visualization and visual analytics. Visual methods effectively communicate the structure, pattern, and trends of variables and their relations. Visualization created within the digital sphere can be used to navigate and query data, enabling users to gain an overview of their data. Visualization may also be used as a form of analytical tool, visual analytics, guided by a combination of algorithms and scientific reasoning which work to extract information, build visual models and explanations, and guide further statistical analysis (Keim et al. 2010). The last but not least challenge is about the stock of descriptive and inferential statistics that have traditionally been used to analyse traditional data. They are also being applied to new data though this is not always straightforward because many of these techniques were developed to draw insights form relatively scarce rather than exhaustive data. Further research is thus required to generate new methods or innovative combination of techniques that can make sense of and extract value from Big Data and data infrastructures.

These challenges are not simply technical, because analytics are the expression of a particular epistemology; therefore, both technical research and epistemological research are required to tackle the challenges of the data revolution. Data revolution is being a great opportunity of innovation of the social sciences. First of all, because it empowers the empirical base of social disciplines, furthermore, because it promotes interdisciplinarity between different areas of science, enhancing integration of data and methods. Only by mixing social theory and computation, data and modelling in an innovative way, social scientists can contribute to a clearer vision of social processes and to the quality of public choices, integrating the more traditional approaches already practiced in social research.

The volume aims to represent the complexity of the whole spectrum of epistemological, technical, and analytical challenges and opportunities that the datafication of society is posing to social sciences. The first section of the volume concentrates on the changes that new data have made to the core of the scientific method and to the theoretical and methodological assumptions that are behind these changes. Moreover, new theoretical reasoning is presented, also on the use of new data for the governance of public policies.

The second section is on methods, software, and data architectures to extract knowledge from data. All the chapters in this section concentrate with the difficult work of data preparation and data curation, focusing on how to manage different forms of data both in structured and in unstructured forms. More specifically, while some contributions deal with the construction of data matrixes ready for statistical analysis, others present softwares or techniques that can help in analysing the different kinds of new data.

The third and the forth parts of the volume present a series of applications. While section three focuses more specifically on the data of the Web (social network data, Web pages and so on), the contributions in section four are applications on data infrastructures' data or data produced by statistical offices. More specifically, in the third section, great relevance is devoted to the techniques such as sentiment analysis, lexical content analysis, and to some innovative efforts to combinate them with social network analysis. The fourth section deals more in depth with the issues of access, integration, and visualization of big databases concentrating both on the analytics required to make sense of them (visualization as well as traditional statistics techniques) and on the techniques needed for their construction.

References

Amaturo, E., & Aragona, B. (2016). La "rivoluzione" dei nuovi dati: quale metodo per il futuro, quale futuro per il metodo?. In F. Corbisiero and E. Ruspini (Eds.), *Sociologia del futuro. Studiare la società del ventunesimo secolo* (pp. 25–50). Milano: CEDAM.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication and Society*, *XV*(5), 662–679. doi:10.1080/1369118X.2012.678878.

Highfield, T., Harrington, S., & Bruns, A. (2013). Twitter as a technology for audiencing and fandom. *Information Communication and Society*, *16*(3), 315–339.

Keim, D., Kohlhammer, J., Ellis, G., & Mansmann, F. (2010). *Mastering the information age solving problems with visual analytics*. Eurographics Association.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, *I*(1), 1–12. doi:10.1177/2053951714528481.

Lazer, D., et al. (2009). Life in the network: The coming age of computational social science. *Science*, *CCCXXIII*(5915), 721–723.

Margetts, H., & Sutcliffe, D. (2013). Addressing the policy challenges and opportunities of "Big data". *Policy & Internet*, *5*(2), 139–146.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work and think*. London: John Murray.

Supiot, A. (2016). *La Gouvernance par les nombres*. Paris: Fayard.

The Economist. (2010). *The data deluge: Businesses, governments and society are only starting to tap its vast potential*, print edition.

Uprichard, E., Burrows, R., & Parker, S. (2009, February 25). Geodemographic code and the production of space. *Environment and Planning A*, *41*(12), 2823–2835.

Part I
Epistemology

On Data, Big Data and Social Research. Is It a Real Revolution?

Federico Neresini

Abstract This chapter aims at discussing critically some epistemological assumptions underlying a data science for social research. For this purpose, it is discussed the general notion of big data and the meaning of key-concepts such as those of information and data, mainly considering contributions coming from the science and technology studies (STS) and the sociology of quantification. In particular, it is argued the necessary shift from a discrete and transportable definition of data to a processual one, also taking into account the fact that data are always a process both when they are produced and when they are used/analysed in order to have research's results. The notion of data-base is compared with that of infrastructure as defined in STS, so that it is clear that they cannot be considered as repositories from which it is possible to extract meanings or results like getting minerals from a mine. Data and data-base are processes which cannot begin without a research question. For these reasons the debate opposing hypothesis-driven versus data-driven research should be overtaken: in social research, as well as in hard sciences, data-driven research simply doesn't exist. The last paragraph is devoted to draw some conclusions from the previous discussion in the form of hopefully useful suggestions for developing a data science for social research.

Keywords Big data • Data-base • Infrastructures • Data-driven/hypothesis driven research • Quantification

Answering the question posed by the title of this contribution might seem easy and straightforward: yes.

In fact it is hard not to recognize that the fast growth of digital data and their increasing availability have opened a new season for social sciences. The unceasing expansion of "datification" or "quantification" (Espeland and Stevens 2008) makes it possible that, for the first time in its history, social research has available a huge amount of data, not only regarding a great variety of phenomena, but also directly

F. Neresini (✉)
FISPPA Department, University of Padua, Padua, Italy
e-mail: federico.neresini@unipd.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_2

and "naturally" generated for the most part by social actors producing those phenomena. The volume of this spontaneous generation of digital data is truly striking: according to some estimates, every minute Google performs 2 million searches and 72 h worth of video is uploaded to YouTube; at the same time there are 1.8 million likes on Facebook, 204 million emails sent and 278,000 tweets posted.[1]

It was hence quite easy to predict that this almost sudden abundance of digital data would attract the interest of many social scientists, as proved by the flourishing of research centres established to exploit this new opportunity and the array of articles in which "big data" are involved.[2]

As a counterbalance to this enthusiasm there have not been lacking—of course, and fortunately—critical reflections, calling attention to the limits of "data-driven" social research (see for example Boyd and Crawford 2012) and to the problems deriving from the quantification processes (see, among others, Espeland and Sauder 2007; Lampland and Star 2009), highlighting the implicit assumptions laying behind the production of digital data by the social media platforms (Gillespie 2014) and the methodological traps to which researchers using, without the necessary awareness, those data and the automatic tools required for handling large amount of digital data are exposed (Giardullo 2015).

It is interesting to note that the debate on big data and social research is proposing again, almost without differences, those arguments that developed at the beginning of the new millennium within the molecular biology research field, and that it is not yet concluded.

The two parties are deployed in two opposite lines: on one side those who are maintaining the so-called "data-first" approach (Golub 2010), and, on the other side, those who are instead affirming the supremacy of the research questions both strategically and operatively orienting the work in the laboratories (Weinberg 2010). This opposition between "data-driven" and "hypothesis-driven" research clearly recalls what was proposed, back in 2008, by Chris Anderson—in a provocative way—as "the end of theory": "With enough data, the numbers speak for themselves" (Anderson 2008).

Already in 2001, John Allen was wondering whether: "With the flood of information from genomics, proteomics, and microarrays, what we really need now is the computer software to tell us what it all means. Or do we?" (Allen 2001). The same question could represent what we are now debating in the case of social sciences; it is enough to substitute the data source: with the flood of information from the web, the official statistics and the record of a huge amount of social activities, what we really need now is the computer software to tell us what it all means. Or do we?

But, this way of addressing the problem, as well as the opposition of data-driven versus hypothesis-driven research and the almost exclusive focus on how to handle

[1] See for example http://blog.qmee.com/qmee-online-in-60-seconds/ (05.06.2016).

[2] Between 2000 and 2015 there were published 2630 articles related to "big data" in the field of social sciences, 1087 only in 2014 and 2015 (Scopus).

data, produce the effect of leaving in the background the fact that data do not exist by themselves, being rather the outcome of a very complex process in which producing and using data are so deeply intertwined that they cannot be considered separately.

Nevertheless, we are inclined to treat distinctly the production of data—i.e. their collection—and the use of them—i.e. their analysis; and this distinction not only induces to paying more attention on the side of data-analysis, but it implicitly suggests also the idea that data simply are there, and that the only problem is how we can use them and with what consequences.

But, first of all, we should not forget that data—no matter how big or small they are—are always the result of a construction process, as should be obvious for social sciences and as is very clear also in the case of the so-called “hard sciences”, at least in the wake of science and technology studies (STS).

Second, using and producing data cannot be considered separately because, on the one hand, the production process affects the possibilities of using data, and, on the other hand, the need to have data to be utilized affects the way they are produced. At the same time, focusing on both sides of producing and using data allows us to pay due attention to what data are, instead of taking them for granted.

Data which populate data-bases available for social sciences today are, in fact, the result of a long and complex process of manufacturing; moreover, the fact that social scientists increasingly seek to use those data for producing new knowledge—together with the fact that these attempts imply a range of problems regarding their accessibility, how to perform queries, the quantity and quality of meta-data, statistical techniques for reducing the complexity associated with their quantity, and the certainly not trivial interpretative work required for making sense of the outputs obtained from data-bases—all of these aspects testify that data entered in a data-base do not live by themselves, but depend on the fact that someone is utilizing them. This is a key point, even if it is very easy to think about “data” as “what remains at the end of these processes”, while, on the contrary, at the end of these processes, nothing remains, because data *are* the process.

1 Data-Bases Are not a Repository

In order to justify the last statement and to explore what it actually implies with regards to the development of a data science for social research, it can be useful to focus our attention on what we still think of as—and therefore still treat as—“bags of data”, i.e. “data-bases”.

The reflection on data-bases has been developed by STS in the field of hard sciences, so that some interesting conclusions they reached can be regarded here as very interesting. It is not by chance that what is going on in the hard sciences can be observed also in the case of the social sciences.

As a starting point, we can refer to this passage by von Foerster, which fits perfectly with the aim of looking at big-data in a critical perspective and, in this case, specifically addressing the relationship between the intrinsic characteristics of what we are used to calling “data” or “information” and their supposed deposits (data-bases):

> Calling these collections of documents ‘information storage and retrieval systems’ is tantamount to calling a garage a ‘transportation storage and retrieval’. By confusing *vehicles* for potential information with information, one puts again the problem of cognition nicely into one’s blind spot of intellectual vision, and the problem conveniently disappears (von Foerster, 1981, p. 237).

So a data-base does not contain data or information, exactly as a garage does not contain transportation, because data, as well as data-bases, are nothing but processes, as has been made very clear by Shannon and Weaver as long ago as 1949:

> Information in communication theory relates not so much to what you do say, as to what you could say. That is, information is a measure of one’s freedom of choice when one selects a message. If one is confronted with a very elementary situation where he has to choose one of two alternative messages, then it is arbitrarily said that the information, associated with this situation, is unity. Note that it is misleading (although often convenient) to say that one or the other message conveys unit information. The concept of information applies not to the individual messages (as the concept of meaning would), but rather to the situation as a whole, the unit information indicating that in this situation one has an amount of freedom of choice, in selecting a message, which it is convenient to regard as a standard or unit amount (Shannon and Weaver, 1949, p. 5).

Hence, precisely as in the case of information, data are not discrete entities, which can be treated as “packages” that can be transmitted from one point to another, or which can be collected and stocked in a deposit, or which can be extracted like precious minerals from a mine. Nevertheless still we substitute the unit of measurement, i.e. a quantity (bit), for what is measured, i.e. the process which, in the case of information, corresponds to reducing uncertainty.

As everybody knows, dimensions actually matter for big-data, supported by a long strain of increasing measures: giga-byte, peta-byte, exa-byte … but very few seem interested in the fact that the unit on which all these measures are based is a process, as clearly stated by Shannon and Weaver. It is possible to find the same conclusion within STS where there is a long array of studies showing the eminently “processual” character of data and data-bases in scientific research and therefore the necessity of not treating data-bases as mere repositories of information. Not only because “raw data is an oxymoron” (Gitelman 2013), but also because data as fixed entities, available for being transferred, transformed or simply used, do not exist. Information—or data—are not discrete elements, well established in time and space, but seamless processes of production and use; outside this process there are no data—nor information.

Being aware of this might lead us to avoid the risk of imperceptible, but—exactly for this reason—very insidious, meaning inversions like that we can see in this passage within a interview by Viktor Mayer-Schoenberg.[3] He maintains that for defining big data we should think about it as follows:

> it's like taking millions of fixed images and mounting them in a movie. The individual fragments, gathered together, take different forms and meanings. This is what happens with the data: the ability to work with a huge amount of numbers allows us to obtain billions of points of view on the world and then to understand it better. Until some time ago it was very expensive and difficult, but new technologies have made these procedures within the reach of many.[4]

We can see here a clear example of the inversion that is the basis on which big data are approached uncritically and naively: data allow us to obtain the points of view, instead of it being the points of view that allow us to generate the data. But a "point of view" is the inescapable starting point of the process which gives rise to data; at the same time, data are not the ending point: they are the process, and therefore we cannot split the expression "processing data" into "processing" and "data" without losing both data and process.

Looking at data which are at stake in doing social research when the data are a huge amount, suggests thinking about a data science for social research as an expression of what has been referred to as "virtual knowledge" and analyzing its relationship to "infrastructure":

> Virtual knowledge is strongly related to the notion that knowledge is embedded in and performed by infrastructures. (…) The infrastructures that are now taking shape are not developed to support well-defined research projects as to the generations of streams of yet undefined research. Most of the data infrastructures that have been built so far have promised the discovery of new patterns and the formation of new-data-driven research. (…) Increasingly, infrastructures and their component network technologies try to support possibility rather than actuality (Wouters, Beaulieu, Scharnhorst, Wyatt, 2013, p. 12).

The concept of "infrastructure", a notion which is clearly and strictly bound up with that of data-base (Mongili and Pellegrino 2015), was introduced into the STS field by Star and Ruhleder almost 20 years ago as follows:

> an infrastructure occurs when local practices are afforded by a larger-scale technology, which can then be used in a natural, ready-to-hand fashion. It becomes transparent as local variations are folded into organizational changes, and becomes an unambiguous home - for somebody. This is not a physical location nor a permanent one, but a working relation - since no home is universal (Star and Ruhleder, 1996, p. 114).

So data-bases, together with all their outfit of standards and routines, are exemplary cases of scientific infrastructures, and also they—as well as data produced, gathered and utilized by them—can exist until they are "in-action".

Moreover, the processual character of data-bases depends also on some aspects intrinsically pertaining to all "things" which can be categorised. It has been pointed

[3] He is the co-author with Cukier of the recently published book "Big-Data: A Revolution That Will Transform How We Live, Work and Think" (Mayer-Schoenberger and Cukier 2012).

[4] La Lettura, Il Corriere della Sera, 01.09.2013, p.14 (our translation).

out by Bowker that many “things” are hard to classify, others do not get classified (i.e. data-bases are selective), others get classified in multiple ways (Bowker 2000).

The data of big data are hence discrete representations of fluid realities—which are actually processes of interaction within a network of heterogeneous actors—they are frames of a film which cannot live outside the film; they appear static and this apparent “staticity” is what makes them exchangeable and transportable, in one word mobile, because they seem detached from the context of their production. For this reason, we should not conceive of data-bases as information’s repositories, not only because data are always generated along a process in which many heterogeneous actors are involved and during which many “translations” occur (Latour 1987, 2005), but also because or, better, mainly because they only exist as processes, and the same goes for the informational infrastructures called data-bases.

2 Some Consequences for Building a Data Science for Social Research

The previous reflections regarding data and data-bases provide the opportunity for pointing out some consequences in order to develop a data science for social research upon fruitful assumptions.

First of all, it is important to stress again the centrality of research questions, and not for abstract reasons related to a supposed supremacy of theory, but mainly because questions play a strategic role in generating data: they create the conditions for facing an uncertainty to be reduced, i.e. for triggering the process through which data are produced and utilized, in both cases through a long array of tools.

Second, we should not forget that those tools—which in the case of big data become digital, as with search engines and their algorithms—are not neutral devices we can decide to use or not. On the one hand we simply cannot have data without this kind of tools; on the other hand, it is not true that “the Internet has no curriculum, no moral values, and no philosophy. It has no religion, no ethnicity, or nationality. It just brings on the data, railroad cars of it, data by the ton” (Sterling 2002, p.51). The Internet only “eclipses intermediaries” (Pariser 2011, p.53). Search engines—Google *in primis*—and other digital tools are not neutral devices, they always offer a selection of the world’s complexity, a selection which is constructed at least for answering in a personalized way needs they ascribe to us as profiled users.

As a third point, the processual character of the digital data with which social research would like to work as well as the un-neutrality of the tools required for retrieving, collecting and processing them make clear what social sciences knew from the very beginning, even if they seem sometimes to forget it: the instruments used for processing data are intrinsically implied in the process of their construction. Put another way, there are not first data, then tools for collecting them, then those for analysing them and finally the results; on the contrary, data which we trust in order to obtain our results depend not only on the questions from which we start, but also on the tools we use for processing them. Exactly as data collected through a questionnaire and analysed with dedicated software are produced both by the questionnaire and by the software, in the same way data pertaining to the social media are produced by the algorithms of the digital platforms on which we "find" them, by the digital tools utilized for processing them and by our research questions, as well as this last depending on the availability of data shaped by the platform and by the tools used for processing them. So yes, questions first, even if questions are not independent from how data are produced and from the tools that can be used.

Furthermore, the heterogeneity of the actors involved in the processes of data construction and data utilization puts forward a very strong argument in favour of the fact that a data science for social sciences cannot be bounded within a single disciplinary domain. As a consequence, an interdisciplinary perspective cannot be avoided, maybe even less so than in the past. But interdisciplinarity is a time-consuming enterprise because it requires a great investment of resources in terms of intermediation among various actors, interests, points of view. It could seem a paradox that in the age of real-time interconnectivity, of fast and easy access to so many digitalized data through the web, of computational power, in short in the era of "speed data", we are requested to be aware of the fact that doing research is a matter of time. It is not by chance that in 2010 many scientists signed the "slow science manifesto": "We do need time to think. We do need time to digest. We do need time to misunderstand each other, especially when fostering lost dialogue between humanities and natural sciences", as is the case of a data science for social research.[5]

It should be clear, therefore, that the necessary interdisciplinarity for a data science even in the field of social research cannot be realized simply by putting together researchers with different training, or proposing training opportunities just as "one near another" classes of different disciplines in a curriculum. Also interdisciplinarity is, in fact, a process which requires time for building it; researchers have to find a new common domain in which they can actually "work together". This process, like any other process, must be fed by motivated actors and must be supported by favourable structural conditions. It means that, for example, it is important to invest in training opportunities for raising a new generation of researchers who have deeply experienced interdisciplinarity, i.e. not offering them just a patchwork of contributions coming from different fields. Moreover, and again as an example, articles published in journals outside the main field of their authors should be recognized institutionally as a valuable contribution and therefore should be considered as relevant in the evaluation exercises devoted to measuring scientific productivity.

[5] http://slow-science.org/ (accessed 16.06.2016).

In other words: building a data science for social research needs not only data and methodological solutions, but also resources, strategically allocated in a long-term strategy of scientific policy which cannot rely only on the goodwill of some social scientists.

References

Allen, J.F. (2001). Bioinformatics and discovery: induction beckons again. *Bioessay*, *23*(1), 104–107.

Anderson, C. (2008). The end of Theory. *WIRED, 16,* 07.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication and Society, 15*(5), 662–679.

Bowker, G. (2000). Biodiversity datadiversity. *Social Studies of Science, 30*(5), 643–684.

Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: how public measures recreate social worlds. *American Journal of Sociology, 113*(1), 1–4.

Espeland, W. N., & Stevens, M. L. (2008). A sociology of quantification. *European Journal of Sociology, 49*(3), 401–437.

Giardullo, P. (2015). Does 'bigger' mean 'better'? Pitfalls and shortcuts associated with big data for social research. *Quality & Quantity*. doi: 10.1007/s11135-015-0162-8.

Gillespie, T. L. (2014) The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, K.A. Foot (Eds.), *Media technologies* (pp. 168–193). Cambridge (MA): MIT Press.

Gitelman, L. (Ed.). (2013). *"Raw data" is an oxymoron*. Cambridge (MA): MIT Press.

Golub, T. (2010). Counterpoint: Data first. *Nature, 464,* 679.

Lampland, M., & Leigh Star, S. (Eds.). (2009). *Standards and their stories. How quantifying, classifying and formalizing practices shape everyday life*. Ithaca: Cornell University Press.

Latour, B. (1987). *Science in action*. Cambridge (MA): Harvard University Press.

Latour, B. (2005). *Re-assembling the social*. Oxford: Oxford University Press.

Mongili, A., & Pellegrino, G. (Eds.). (2015). *Information infrastructure(s): Boundaries, ecologies, multiplicity*. Newcastle: Cambridge Scholars Publishing.

Pariser, E. (2011). *The filter bubble. What the internet is hiding from you*. New York: Penguin Books.

Mayer-Schoenberger, V. & Cukier K. (2012). *Big data: A revolution that will transform how we live, work and think*. New York: Houghton Mifflin.

Shannon, C. H., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research, 7*(1), 111–134.

Sterling, B. (2002). *Tomorrow now. Envisioning the next fifty years*. New York: Random House Inc.

von Foerster, H. (1981). *Observing systems*. Seaside: Intersystems Publications.

Weinberg, R. A. (2010). Point: Hypothesis first. *Nature, 464,* 678.

Wouters, P., Beaulieu, A., Scharnhorst, A., & Wyatt, S. (Eds.). (2013). *Virtual knowledge. Experimenting in the humanities and the social sciences*. Cambridge (MA): MIT Press.

New Data Science: The Sociological Point of View

Biagio Aragona

Abstract The objective of this chapter is to introduce the contribution that, apart from post-positivism, other sociological paradigms, such as interpretivism and social constructionism, may give to the development of research and thinking about data science in social research. These two paradigms have theoretical and methodological beliefs that seem unfitted to interpret the data revolution era because they are focused on individuals, *verstehen* and sense of action and have been usually associated with qualitative research. But they may be of great help in addressing the future of new data research in our discipline, especially on two important aspects: first, on how objective new data are and, furthermore, on the role of knowledge in new data use and construction.

Keywords New data • Data culture • Data assemblage • Social constructionism • Post-positivism

1 The Reframing of Social Sciences

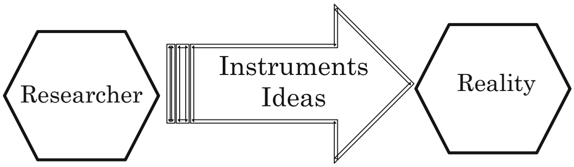
Changes in social networking and the pervasive and ubiquitous Web use in daily life, as well as improvements in computational power and data storage, are having impressive effects on data production and usage. Social networks, sensors and data infrastructure are generating a massive amount of new data (big data, big corpora, linked data, open data, etc.) that are readily available for the analysis of societies.

A first consequence of the passage from data scarce to data-intensive societies has been the re-emergence of data-driven science, which is opposed to theory-laden science that is typical of post-positivist social sciences. This form of neo-empiricism implies both an epistemological and an ontological assumption. First of all, it sustains the adherence to inductivism, where a proposition is scientific

B. Aragona (✉)
Department of Social Sciences, University of Naples Federico II,
Vico Monte Di Pietà 1, Naples, Italy
e-mail: aragona@unina.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_3

Fig. 1 Data collection in post-positivist view



if proven from facts, so research is data-driven. Furthermore, it supports data objectivity, where data are considered as neutral observation of reality; they are reality.

Neo-positivism and post-positivism, already in the mid of last century, have criticized both inductivism and data objectivity. Popper (1967, 1972) believed that no proposition can be proven from facts and that hypotheses are always theory-laden. Lackatos (1976) regarded empirical support as a three-place relationship between theory, evidence and background knowledge, where the latter is represented by the whole set of facts and parameters used in the construction of any given theory and, it could be added, of any given data.

According to post-positivist thinking, data are not objective and neutral. What we have is the researcher, the reality and, between these two, instruments (methods and techniques) and ideas (theories and empirical hypotheses) (Fig. 1).

As well as neo- and post-positivism started to criticize the ingénue epistemic and methodological view of first positivism, critical data studies started to criticize the neo-empiricism brought about by the present data revolution. Dalton and Tatcher (2014) have called critical data studies those studies that apply critical social theory to data to explore the ways in which they are never neutral, objective, raw representation of the world but are situated, contingent, relational and contextual. In this context, data are considered as complex socio-technical systems that are embedded within a larger institutional landscape of researchers, institutions and corporations (Ruppert 2013). And the inductive method as in the Anderson vision of the “end of theory” (2008) is considered an unsupportable fantasy. For Kitchin, for example, more common is the use of abduction, which enables to fit unexpected findings into an interpretative framework (Kitchin 2014). Abduction is also called the inference of best explanation, and it is for example applied through the Bayesian method of explanation called maximum likelihood. Peirce (1883) gave its first formulation, but for him abduction was typical of the context of discovery, while modern epistemology attributes it to the context of justification.

A further epistemological consequence of data revolution has been a redefinition of the boundaries of disciplines and the foundation of new interdisciplinary fields where technology plays a greater role, such as computational social science and digital humanities. In sociology, data revolution is undermining the already weak boundaries between quantitative and qualitative research. The big data realm in fact blends differences between textual and numerical data. Often numbers, texts and even images merge into the same database. Moreover, user-generated contents on social networks and Websites, which constitute a massive amount of data, are classified and analysed through techniques of sense making and meaning construction that have the features of deep data analytics, which are typical of qualitative research (Boccia Artieri 2015). That is why *mixed methods* (Hesse-Biber and Johnson 2013) researchers, who have always been keen to integration between quality and quantity, are paying great attention to the developments of new data use in sociology.

2 Data as Signification Acts

One difference between the epistemic position of Max Weber and that of neo-positivists and post-positivists is in the role that the point of view has in the connection between data, reality and knowledge. For Lackatos (1976), facts are just facts (data are just data), while the interpretation of facts (data) depends upon economic interests and points of view. For Weber, not only the interpretation of facts depends upon point of views, but also facts depend on point of views. As he has firstly stated (1904), the role of researchers in the construction of data is high and the distinguishing criteria used in their capture have consequences on the results.

Positivist and interpretivist paradigms have been put together inside social constructionism by Berger and Luckmann, referring to the work of Alfred Schutz (Schutz 1960; Berger and Luckmann 1966). It is interesting how Berger and Luckmann support the objectivity of data by defining them as signification acts. Signification is an objectivation through the definition of a sign: “able to communicate meanings that are not directs manifestations of *hic et nunc* subjectivity” (Berger and Luckmann 1966/1969, 58). The most important signification acts in history are language and numeration, actually also the two main forms of data. In a context of high data production as the contemporary societies that definition is meaningful, because it points out the symbolic dimension of both social life and data construction.

It is quite surprising how a mathematician as Tobias Dantzig shared the same notion of data as signification acts (1930/1954). He writes:

> Signification allows to transcend the subjective reality and pass to the objective reality. That reality is not a collection of frozen images, but a living, growing organism(…) an individual without a milieu, deprived of language, deprived of all opportunity to exchange impressions with his peers, could **not construct a science of number (a data science**[1]**)**. To his perceptual world data would have no reality, no meaning. (Dantzig 1930/1954, 253).

Both Dantzig and the constructionists hence recall the cognitive and symbolic aspects of data and believe that simple facts do not exist, but “facts are always interpreted” (Schutz 1962, 5). What Schutz and their fellows have proposed is a methodological constructionism where there is a suspension of ontology. Objectivity of data does not arise from a supposed reality but from the agreement of all

[1] Bold text and text in brackets are added by the author.

observers and through procedures which are intersubjectively defined. Scientific knowledge is therefore grounded on the agreement between observers, and it is not independent from the questions whose answers have been given.

3 Data Culture, Data Assemblage and Data Science

Two are the main reasons why social constructionism seems fitted to cross the data revolution era: first of all, because it is focused on daily life and new data bring in many ways the daily life of individuals in the hearth of social knowledge; furthermore, because it believes that the empirical process and the construction of reality are based upon the agreement of observers. And this perspective is useful to address the future of social sciences and data science into an interdisciplinary context.

Ubiquitous and pervasive technology brings daily life in data production. Social networks data, smartphone data and user-generated contents on the Internet are windows upon the subjectivity of individuals. These new data are able to track, trace, record and sense our complex interactions with the social world. For example, one point about the supremacy of big data on traditional data is that researchers may be in the action instead of collecting account of actions (through survey) (Savage and Burrows 2014). But the fact that there is no contact between researcher and subjects causes advantages as well as disadvantages. For example, automated and volunteer data are just founded data produced in an unobtrusive way (Webb et al. 1966). Unobtrusive methods collect data on research units that are not aware of being studied. There is not active participation (feedback) from those being researched; they are not-reactive. The most important advantage of using not-reactive methods is that the researcher does not perturb the behaviours of the subjects he/she is studying. There is no reaction to questions posed or observations made. The experimental "Hawthorne effect" (Mayo 1933) and the answers' reliability problems (social desirability, memory effects, acquiescence) just vanish.

On the contrary, it is more important to study what is the context of validity of these data. Special attention must be paid to user-generated contents on social network and Websites. They represent a massive quantity of data, but their contents, and the ways of making sense of them through classifications, have the characteristics of deep data, those that are used in the realm of qualitative sociology. (Boccia Artieri 2015).

From a sociological perspective, changes in data construction are also changing the role of social actors in the production of data and their data culture. Data culture refers to the connection between the moment when data are constructed and the moment when they are used to produce knowledge in a specific domain (Sgritta 1988; Aragona 2008). Data culture is defined by two elements: the organizational and methodological changes which constitute the production of data in a specific time and the quantity of social data. The actual data culture era sees an impressive amount of social data produced daily, where the use of technology is more than ever. Producers and users may be very distant, and data that are generated by someone may be shared, sold, combined, merged and then analysed to produce knowledge on some specific domains. In this context, where many actors are involved in the production and use of data, empirical process truly becomes a cultural process which needs to be understood as such. What all this view on data suggests is that to put meaning into data and to understand what piece of reality that data is representing, we need to have a close look on what Kitchin and Lauriault (2014) call data assemblages.

Data assemblages are complex socio-technical system composed of many apparatuses and elements that are thoroughly entwined, whose central concern is the production of data. Data assemblages are made of two main activities:

- a technical process (operational definitions, data selection, data curation) which shapes the data as they are;
- a cultural process, which shapes the background knowledge (believes, instruments and others things that are shared in a scientific community) which enables the sharing of meanings.

Data science in social research requires therefore interdisciplinary and cross-cutting approaches, combining skills and viewpoints that cut across disciplines. The expertises needed are domain expertise, data expertise and analytical expertise. A dialogue on different technical and cultural processes is required to blend methodologies and disciplinary matrixes and shape what Lackatos (1976) called background knowledge (the whole set of facts and parameters used in the construction of any given theory and of any given data).

Methodological constructionism may be the right approach for addressing the future of new data research in social sciences.

4 Sense Making, Small Decisions and Social Actors

Some examples drawn from empirical research may clarify the contribution that constructivism may give to the development of new data research in social sciences.

A first example is the construction of meanings that is used to perform data curation in the analysis of *big corpora*, when researcher transforms unstructured texts in databases which can be analyzed through computational techniques. Di Maggio (2015) notes that topic-modelling programs require lots of decisions that most social scientists are ill-equipped to make. For example, it is essential to document the crawler's specification in detail to meet the standards of peer review and replicability, but not all social scientists are able to understand crawler language. Most textual analysis run on Web data requires close reading that has traditionally been conducted by hermeneutically oriented scholars who find not one simple uncontested communication, but multiple, contradictory and overlapping meanings.

Another point is the role of small decisions on research results. Jana Diesner (2015) states that small decisions concerning data construction and data preparation, decisions that are often not given careful attention and about which there are few or no "best practices", can have enormous (often undesired) impact on the results of new data analysis. What it seems dangerous is if the role of small decisions in data construction is deliberately hidden with the scope to fasten what Marradi (1994) calls the "protective belt of objectivity". Acknowledging the role of small decisions in fact would crumble the positivist vision of data as pictures of reality. In practice, however, research follows a series of choices made by the researcher. It should therefore be possible to understand the kind of choices that he/she made and to inspect the empirical process. As Marradi ironically notes, considering the researcher as a photographer and empirical observations as pictures of reality, so that the only decision to construct data is making the picture, seems reductive even for the activity of a photographer. The point of view is, as Weber said, something that counts throughout the empirical process, and it must be explicated.

Other examples that show the need to approach new data research (or data science) in a constructionist way are about the role that social actors may have in promoting certain images of the social world. Adams and Bruckner (2016), for instance, show that the social system operating at the heart of the Wikipedia's Website generates biases in the production process and in the database itself. A site that should promote democracy clearly shows that decisions about content and form are now made top-down and not bottom-up and that new users may encounter a hostile environment that is not always open to their contributions (Schneider et al. 2014). Researchers who make use of Wikipedia should be aware of that and keep it in account when giving meaning to their wiki-data. Furthermore, Taylor et al. (2014) note that the access to corporate data in economics may address research results, because it is proprietary and tends to be offered to researchers only subject to non-disclosure agreements which may limit the replicability of studies.

Beside that, new data rise even more critical sociological questions about to what extent these data may be used for the analysis of society and how they can be effective for designing, implementing, monitoring and evaluating social policies. This latter point is crucial to understand what is the future of new data within both sociology and society. In the framework of evidence-based policies, data as a form of "governing knowledge" has become a key focus in studies of social policy at national and global scales, but little research has focused on how new digital data facilitate governments to produce such policies. This is despite the fact that digital data have been "normalized" within administrations, as evidenced by proliferating database-related technologies of governance and by the work that official statistic is doing for implementing new data in their production processes.

5 Final Remarks

In interdisciplinary contexts, a constructionist approach may be useful to make explicit the interpretative schemes that researchers use to make data objective. Data are a construction of the observer, and they are embedded to a specific point of view which must be explicated in all its components.

The agreement between observers is a common ground on which different disciplines may gather. It has been already mentioned the vision of the mathematician Dantzig and beside him, statisticians as Desroisieres (1998) has for example noted that data are not neutral, they are political. And economists such as Sen have highlighted that selection and definition of what has to become the informational basis of judgement (Sen 1982) is the result of compromises between many actors who adopt different cognitive patterns.

The problem of the connection between reality, data and theory is more important than ever, because many actors with different cultural and technical backgrounds are involved in the production and use of new data, and they all aliment the data assemblages. The question is then what is the role of actors' interpretations in shaping the data as they are?

Sociologists should therefore develop more research about not only the technical addresses which guide new data research, but also about the normative and political imperatives that shape the background knowledge that it is used to make sense of new data. It is because they, more than others, have in their epistemological traditions critical approaches to data, reality and social knowledge. Maybe, after the computational turn, it is now time for the constructionist turn.

References

Adams, J., & Bruckner, C. S. (2016). Wikipedia, sociology and the promise and pitfalls of big data. *Big Data and Society, 2*(2), 1–5. doi:10.1177/20253951715607482.

Anderson, C. (2008, June 23). The end of theory: The data deluge makes the scientific method obsolete. *Wired*. Retrieved Jan 09, 2016 from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

Aragona, B. (2008). Una nuova cultura del dato. *Sociologia e ricerca sociale, XXIX, 87,* 159–172. doi:10.3280/SR2008-087004.

Berger, P., & Luckmann, T. (1966). The social construction of reality. tr.it. La realtà come costruzione sociale 1969 Il Mulino.

Boccia Artieri, G. (2015). *Gli effetti sociali del web. Forme della comunicazione e metodologie della ricerca online*. Milano: FrancoAngeli.

Dalton, C. & Thatcher, J. (2014). What does a critical data studies look like, and why do we care? Seven points for a critical approach to 'big data'. *Society and Space open site*.

Dantzig, T. (1930). *Number, the language of science*. New York: Scribner. Desroiséres, A. (1998). *The politics of large numbers: a history of statistical reasoning*. Cambridge: Harward University Press.

Diesner, J. (2015). Small decisions with big impact on data analytics. *Big Data & Society, 2*(2), 24–32. doi:10.1177/20253951715617185.

Di Maggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society, 2*(2), 6–14. doi:10.1177/20253951715602908.

Hesse-Biber, S. N., & Johnson, R. B. (2013). Coming at things differently future directions of possible engagement with mixed methods research. *Journal of Mixed Methods Research, VII, 2,* 103–109. doi:10.1177/1558689813483987.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1), 1–12. doi:10.1177/2053951714528481.

Kitchin, R., Lauriault, T. P. (2014). Towards critical data studies: Charting and unpacking data assemblages and their work. *The programmable city working paper*, 2.

Lackatos, I. (1976). Proof and refutations. In J. Worral and E. Zahar (Eds.), *The logic of mathematical discovery*. Cambridge: Cambridge University Press.

Mayo, E. (1933). *The human problems of an industrial civilisation*. New York: MacMillan.

Marradi, A. (1994). Referenti, pensiero e linguaggio: una questione rilevante per gli indicatori. *Sociologia e ricerca sociale, XV, 43,* 137–207.

Peirce, C. S. (1883). (Ed.), *Studies in logic*. Boston, MA: Little, Brown and Company.

Popper, K. R. (1967). Knowledge: Subjective versus objective. In D. Miller (Ed.), 1983 *A Pocket Popper* (pp. 58–77). Oxford: Oxford University Press.

Popper, K. R. (1972). *Objective knowledge. An evolutionary approach*. Oxford: Clarendon Press.

Ruppert, E. (2013). Rethinking empirical social sciences. *Dialogues in Human Geography, 3*(3), 268–273.

Savage, M., & Burrows, R. (2014). After the crisis? Big data and the methodological challenges of empirical sociology. *Big Data and Society*, April-June, 1–6. doi:10.1177/2053951714540280.

Schneider, J., Gelley, B. S., & Halfaker, A. (2014). Accept, decline, postpone: How newcomer productivity is reduced in English Wikipedia by pre-publication review. In *Proceedings of the international symposium on open collaboration*, Berlin, Germany, 27–29 August.

Schutz, A. (1960). The social world and the theory of social action. *Social Research*, 203–221.

Schutz, A. (1962). *Common sense and scientific interpretation of human action in collected papers I* (pp. 3–47). Heidelberg: Springer.

Sen, A. (1982). *Choice, welfare and measurement*. Cambridge, MA: Harward University Press.

Sgritta, G. (1988). Introduzione. In L. Benvenuti, *Immagini della società italiana*, Roma, Istat.

Taylor, L., Schroeder, R., & Meyer, E. (2014). Emerging practices and perspectives on big data analysis in economics: Bigger and better or more of the same? *Big Data & Society, 1*(2), 7–16. doi:10.1177/20253951715602908.

Webb, E. J., Campbell, D. T., & Schwartz, R. D. (1966). *Unobtrusive methods: Non-reactive research in the social sciences*. Chicago: Rand McNally.

Weber, M. (1904). *Die « Objektivität » sozialwissenschaftlicher und sozialpolitischer Erkenntnis*. XIX: Archiv für Sozialwissenschaft und Sozialpolitik.

Data Revolutions in Sociology

Barbara Saracino

Abstract Following the French Revolution, society becomes the leading dimension of existence, more prominent than the state and government. According to the "hygienic party", who promoted health and education, statistics can contribute to an understanding not only of the main causes of disease and death, but also of crime and unrest; scientific ground can be provided for social policy. All kinds of social phenomena are classified, measured, and made public, and large-scale statistical recurrences are identified. In the "avalanche of numbers" published, the first sociologists recognized mass regularities in order to identify laws of human behaviour. Nowadays, the availability of Big Data paves the way to new epistemological challenges. Will the new data revolution lead to a new paradigm in Sociology?

Keywords History of statistics • Big Data • Data deluge • Computational turn • Sociological knowledge

"The era of Big Data has begun. Computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and other scholars are clamouring for access to the massive quantities of information produced by and about people, things, and their interactions. Diverse groups argue about the potential benefits and costs of analysing genetic sequences, social media interactions, health records, phone logs, government records, and other digital traces left by people. Significant questions emerge" (Boyd and Crawford 2012, 662).

The purpose of this paper is to analyse the historical period that witnessed the rise of social mechanics and the birth of sociology as specialized discipline and thus to contribute to the reflection on possible present and future developments, consequences, and implications of the use of Big Data for the production and development of sociological knowledge.

B. Saracino (✉)
Department of Social Sciences, University of Naples Federico II,
Vico Monte Della Pietà 1, Naples, Italy
e-mail: barbara.saracino@unina.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_4

In the seventeenth century, the idea that social phenomena could undergo quantitative analysis gained momentum. Demographic problems were the first to be discussed systematically, as the wide-spreading insurance systems required accurate numerical basis, and the size of the population was considered a crucial element for state power and wealth.

The study of social statistics started in 1660, and since that time for around a century and a half, it was known as "political arithmetic". Its purpose was the promotion of upright and well-documented state policies. In his *Observations upon the Bills of Mortality,* John Graunt wrote: "That whereas the Art of Governing, and the true Politicks, is how to preserve the Subject in Peace and Plenty; that men study only that part of it which teacheth how to supplant and reach one another, and how, not by fair out-running but tripping up each other's heels, to win the Prize. Now, the Foundation or Elements of this honest harmless Policy is to understand the Land, and the Hands of the Territory, to be governed according to all their intrinsic and accidental differences" (Porter 1986, 18). Policies had to rest on a profound understanding of the territory and its inhabitants, on concrete knowledge expressed in terms of numbers, weights, and measures.

According to Hacking (1990), Graunt and the English started the public use of statistical data, and Italian philosophers were the creators of the modern notion of state, but it was German thinkers and statesmen who were the first to gain awareness of the importance of data collecting. Instead of leaving it to personal initiative, a nation state needed to establish a specific organization, a central statistical office, in charge of collecting all the data necessary to define its own size and power. Leibniz was the spiritual father of Prussian official statistics. In 1685, he affirmed that a Prussian state had to be established, that the measure of the power of a state was its population, and that a state needed a central statistical office in order to know its own power.

According to Leibniz, this office had to be at disposal of all the administration branches and had the task to maintain a central register of deaths, baptisms, and marriages, in order to allow the estimation of population size. In those days, a general population census was deemed impracticable, as the numerousness of the population of a country, unlike a walled city or a colony, was not a measurable quantity. Only the establishment of designated institutions would finally allow it.

Leibniz was extremely interested in statistical questions of all sorts and pursued a rich correspondence about many issues of public health and demography. Prince Frederick II of Prussia wanted to be king of a united Brandenburg and Prussia, and Leibniz urged his case. Frederick's opponents argued that Prussia could provide only a limited contribution to unification with Brandenburg, so that the king should not be Prussian. But that was a mistake, according to Leibniz, as the real measure of the power of a kingdom was the number of its subjects: 65.400 children were born every year in the whole region, and 22.680 were Prussian, so Prussia was vital. Leibniz wrote these notes in 1700; the following year, the kingdom of Brandenburg-Prussia was established. Some years later, court officers created a system to register births, deaths, and weddings in the four main cities of the kingdom. In 1733, the data about the population became a state secret.[1] During the Seven Years' War, a third of the population was decimated and colonization was required in order to restore ravished farmland. During the reign of Frederick, the list of the things that were counted extended up to seven pages.

Beyond state officers and private citizens collecting demographic data, such as Süssmilch, and beyond political arithmetic, Germany witnessed the development of "university statistics". The work of university statisticians was almost never quantitative. They feared that the indiscriminate use of data would add a materialistic character to the comparative study of states, ending up undermining educational and social value of their teaching. University statisticians believed it was necessary to distinguish between two sciences: a descriptive and non-numerical science, which was theirs, and another science that was heir of English political arithmetic (Lazarsfeld 1961).

At the beginning of nineteenth century, in Great Britain and France, political arithmetic was replaced by statistics. The change was not only terminological, but reflected a substantive transformation. Numerical statistics inherited an extraordinarily large field of application, from geography to climate, from commerce to population and culture. Statisticians started to investigate about all kinds of institutions and to collect data about commerce, industrial progress, work, poverty, education, health care, and crime. The extension of the field of numerical surveys is combined with an important change in the conception of their purpose. That may appear clearer through the comparison between two famous scholars, Süssmilch and Malthus, respectively, before and after the French Revolution (Porter 1986).

Before the French Revolution, Süssmilch, starting from the premise that population growth was the main aim of every ruler, devoted his work to show what the prince could do to promote demographic increase. After the French Revolution, Malthus argued that high density of population was the major cause of misery and poor health in a country. Population was not something flexible that could be manipulated, but the product of persistent customs and natural laws. Government could not dominate society, for it was itself conditioned by it. Malthus believed that society was a dynamic and potentially instable force, a source of trouble. Through statistical surveys, political leaders could have the chance to know the people and attempt to avoid disorders, introducing public education and informing about the true causes of poverty.

Collini (1980, 203–204) has written that in the first half of the nineteenth century, European intellectuals started to consider the dimension of society more central than state and government. Society was seen both as source of progress, constituting labour force for industrialization, and as cause of instability, symbolized by the French Revolution and by the incessant troubles in all Europe. "The emphasis upon the priority of the social came to be closely bound up with two characteristic features of nineteenth-century thought in general. The first was a

[1] Statistical data were treated as strictly confidential due to their potential military value. Later, revolutionary governments would proclaim the necessity to publicize them.

widely ramifying historicism [...] The second was a profound commitment to a conception of the methods of natural science as man's only reliable cognitive relation to the world, and hence as the model for the study of human behaviour. Taken together, these beliefs constituted a character for a science of society, the project of discovering the natural laws which governed social development, and upon which political prescription and action were alike dependent".

During the nineteenth century, a growing number of scientists began to search for mass regularities and to overlook the causes of single events. They gradually realized with astonishment that single disordered, chaotic, or irrational phenomena showed unexpected regularities on large scale, and thus, they established a new type of law, the statistical law.[2] According to Hacking (1990), to ascertain the existence of statistical laws, both observation of regularities on large scale and a "right kind of readers" were necessary. Regularities became visible when social phenomena were classified, quantified, and publicized, that is to say after the "avalanche of numbers" published at the beginning of the nineteenth century. The right kind of readers, ready to find analogies between laws of society and laws of nature, were Western European intellectuals.[3]

The period defined by Westergaard (1932) as the "era of enthusiasm" for statistics started in the first decades of the nineteenth century in France and then developed with the Victorian statistical movement.

As Coleman (1982) has pointed out, starting from the 1820s in France, some "defenders of public health", especially military doctors on retirement after Napoleonic wars, took the initiative and conducted quantitative investigations. Their general interests were health and education, and collecting data could help them to understand causes of disease and death, criminality, and revolt. It seemed thus possible to obtain a scientific basis for a social policy.

Cullen (1975) has written that research in England, from the first half of Victorian age, had the main scope to celebrate industrial progress, blaming other causes, such as alcohol, moral degradation, and urbanization, for social unrest. Ignorance and dirt were considered responsible for diffusion of diseases, growth of criminality, and risk of national disorder within the working class. Statistical investigation would provide the empirical support for the necessary reforms.

According to Funkhouser (1937, 291), "an interesting development in the history of statistics is that of the gradual merging of political arithmetic and the theory of probabilities into a science of statistics in the first part of the nineteenth century. The students of political arithmetic had the urge for the scientific study of anthropological and political questions and were slowly improving their data in

[2] In 1889, Galton affirmed that the law of error "reigns with serenity and in complete self-effacement amidst the wildest confusion. The huger the mob and the greater the apparent anarchy, the more perfect is its sway".

[3] In *The Taming of Chance,* Hacking (1990) compared Prussian (and Eastern European) attitude towards numerical data with the position of scholars in Great Britain, France, and the other countries of Western Europe. It was the West, where libertarian, individualistic, and atomistic notions of person and state were spreading, to start to formulate social laws based on data.

quantity and quality, but they lacked a powerful enough tool to handle their problems. This tool was provided in the theory of probabilities". One of the most committed supporters of the application of probability theory to the study of social phenomena was Adolphe Quetelet.

Influenced by the works of Laplace and Fourier, Quetelet started to believe, around 1830, in the possibility of applying the methods of physical and natural sciences to human activities, going beyond the mere collection of data that was so fashionable at the time. Therefore, he starts to dig out statistical laws from the avalanche of published data, finds the cause of the observed demographic regularities in forces acting within society, and combines statistical interests with astronomical and mathematical instruments. In his opinion, mathematics can bring order out of the apparent social chaos and can give the chance to dominate scientifically seemingly uncontrolled social phenomena. The application of the law of errors to the distribution of human characteristics permits to confirm the hypothesis of social physics and to demonstrate that concepts and instruments of astronomy are the most adequate to catch the essential characters of the human being, the only entity that, until that time, was deemed impermeable to science. Quetelet thinks that persons are imperfect copies of the average man and that their growth is influenced by a large series of accidental causes and errors, just like the exact observation of an astronomical object or event. The function of the errors is useful to define a "type" and to identify that single cause of phenomena that is usually obscured by the action of disturbing causes. The first positivist sociologists considered Queletet's researches about humankind as a valid support to their idea of normality. The Belgian author conceived normality as optimal status to achieve; his idea was echoed by Durkheim who went as far as to oppose normal to pathological status, meant as deviation from the norm.

Nearly two centuries after the era of enthusiasm for statistics and Quetelet's works, we are discussing the pervasive character of the phenomenon of the so-called datification. The debate concerns the proliferation of data and information in the contemporary society of knowledge—founded on the diffusion of digital society, computers, informatics culture, and importance of Internet and made possible thanks to "information infrastructures" such as databases, networks, and interfaces (Lorenzet 2015)—but is also about the growing importance of quantification processes in organization contexts and of "governance by numbers" (Supiot 2015).

Due to the ongoing increasing digitalization, we are now witnessing a process of acceleration that is causing not only quantitative but also qualitative effects on the ways to realize and produce knowledge. One of the most distinctive traits of the current attention dedicated to Big Data[4] is perhaps the growing trust in machines for

[4] With Big Data, I intend here extensively all the data and information that are capable of meeting the requirements of all the characteristics that have been identified in the literature, such as the three "Vs": Volume, Velocity, and Variety (Giardullo 2015, 2).

the production of knowledge, combined with the shift from prevailing mechanical–analog technology to digital–algorithmic devices (Neresini 2015).

“To mediate an object, a digital or computational device requires that this object be translated into the digital code that it can understand. This minimal transformation is effected through the input mechanism of a socio-technical device within which a model or image is stabilised and attended to. It is then internally transformed, depending on a number of interventions, processes or filters, and eventually displayed as a final calculation, usually in a visual form. [...] In other words, a computer requires that everything is transformed from the continuous flow of our everyday reality into a grid of numbers that can be stored as a representation of reality which can then be manipulated using algorithms” (Berry 2011, 1–2).

Digital devices have enabled a strong acceleration of the tendency towards mathematization which characterizes modern sciences and thus the way in which we describe, analyse, and intervene in reality to modify it. The so-called computational turn has enormously amplified the importance of one of the fundamental principles of the possibility of accumulating and producing knowledge, that is to say the capability of “acting at a distance”.

According to Latour (1987), it is not possible to describe knowledge in itself, by opposing it to ignorance or to belief, as its very sense raises only when taking into consideration an entire cycle of accumulation. To accumulate means to acquire a familiarity with distant things, events, and people. How can you act at a distance on unfamiliar objects? By bringing them home, somehow. How can you do it, given that they are distant? By inventing devices that render them mobile, combinable, and stable. This mixture of mobility and combinability permits to dominate at a distance. Inscriptions (formulas, tables, and charts) accelerate the accumulation movement. To dominate at a distance, several actions are necessary: firstly, to translate the world in order to make it enter in “centres of calculation”; secondly, many elements must be moved from a distance without being really introduced inside to avoid their flooding in centres of calculation; and thirdly, new codes must be invented to hold maximum information in minimum space. Operating on the centres through a series of subsequent representations permits to obtain and keep an advantage: it is possible to obtain representations of nth-order that are combinable with other representations of nth-order giving the same mathematical structure to every element. The process of abstraction thus enables the gathering of the maximum amount of information in one single place.

If nothing is more stable, mobile, and combinable of a number in digital format, digitalization has brought to the utmost extreme the processes of abstraction, standardization, and action at a distance, with the double effect both of generating big quantities of data in numerical form and of increasing their agency.[5] The

[5] “Perhaps the most important element distinguishing Big Data from other huge collections of data, that is, census data, is the fast and automatic generation of a high volume of information, which means delegating data collection to an automatic device. In fact, huge databases are ‘populated’ through specific scripts that are nested in servers and types of counter machinery” (Giardullo 2015, 2).

production and the treatment of big quantities of data have made the role of algorithms everyday more necessary in the process of construction of knowledge.

Technically speaking, an algorithm is a codified procedure to transform an input into an output, "but as we have embraced computational tools as our primary media of expression, and have made all information digital, we are subjecting human discourse and knowledge to these procedural logics that undergird all computation. And there are specific implications when we use algorithms to select what is most relevant from a corpus of data composed of traces of our activities, preferences, and expressions". Gillispie (2014, 167–168) has considered algorithms that manage information not only as codes lines but also as "new knowledge logic", and he has identified different characteristics of their unprecedented "public relevance". Their public relevance includes the choice about what to include or exclude in the preparation of data; the implications of the attempts to know and predict algorithms' users; the criteria by which what is relevant is determined; the "algorithmic objectivity", that is to say how to position the algorithm in the face of controversy as an assurance of impartiality due to its technical character; the reshaping of users' practices in response to the algorithms they depend on; and the production of "calculated publics".

The expression "data deluge" was created within the Human Genome Project at the beginning of the 1990s to indicate the enormous quantity of data that molecular biology was starting to deal with. In most recent years, data deluge is overwhelming even the work of social scientists.

The increasing availability of Big Data for research (in private or public institutions) and the design of interventions (from marketing to public administration) have divided researchers into two opposite sides: on the one hand, the sceptics, who fundamentally question the legitimacy of the use of such data on the basis of privacy issues and other ethical concerns and on the other hand, the enthusiasts, who focus on the transformational impact of having more information than ever before (Gonzáles-Bailón 2013, 148).

As argued in *Nature* (2007, 637–638), "For a certain sort of social scientist, the traffic patterns of millions of e-mails look like manna from heaven. Such data sets allow them to map formal and informal networks and pecking orders, to see how interactions affect an organization's function, and to watch these elements evolve over time. They are emblematic of the vast amounts of structured information opening up new ways to study communities and societies. Such research could provide much-needed insight into some of the most pressing issues of our day, from the functioning of religious fundamentalism to the way behaviour influences epidemics […] But for such research to flourish, it must engender that which it seeks to describe […] Any data on human subjects inevitably raise privacy issues, and the real risks of abuse of such data are difficult to quantify".

However, even the enthusiasts have divided into two groups. As the availability of big quantities of data has grown, researchers have engaged in a debate about the opposition between hypothesis-driven research and data-driven research. They first believe that Big Data will radically change the way in which we make sense of the world: the data speak for themselves, and theoretical interpretative models are not necessary. The main argument of those who proclaim the "end of theory" (Anderson 2008) is that the most measured and recorded age in history demands a different approach to data: in other words, being able to track human behaviour with unprecedented fidelity and precision is more powerful than imperfect models of why people behave the way they do. Alternatively, there are those who believe the exact opposite: that theory and interpretation are more necessary than ever before if we are to find the appropriate layer of information, to disentangle signal from noise, to identify meaningful correlations, and to discard those that are unsubstantial (Gonzáles-Bailón 2013, 148).

In the "avalanche of numbers" published at the beginning of the nineteenth century, the first sociologists recognized mass regularities in order to identify laws of human behaviour. Nowadays, the availability of Big Data paves the way to new epistemological challenges. Will the new data revolution lead to a new paradigm in Sociology?

References

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*.

Berry, D. M. (2011). The computational turn. Thinking about the digital humanities. *Culture Machine, 12,* 1–22.

Boyd, D. & Crawford, K. (2012). Critical questions for big data. *Information Communication & Society*, 15(5), 662–679.

Coleman, W. (1982). *Death is a social disease. public health and political economy in early industrial france*. Madison: The University of Wisconsin Press.

Collini, S. (1980). Political theory and the 'science of society' in Victorian Britain. *The Historical Journal, 23*(1), 203–231.

Cullen, M. (1975). *The statistical movement in early Victorian Britain: The foundation of empirical social research*. Hassocks: Harvester.

Funkhouser, H. G. (1937). Historical development of the graphical representation of statistical data. *Osiris, 3,* 269–404.

Galton, F. (1889). *Natural inheritance*. London: Macmillan.

Giardullo, P. (2015). Does "bigger" mean "better"? Pitfalls and shortcuts associated with big data for social research. *Quality and Quantity*, 1–19.

Gillespie, T. (2014). The relevance of algorithms. In T. Gillispie, P. Broczkowski, & K. Foot (Eds.), *Media technologies: Essays on communication, materiality and society* (pp. 167–194). Cambridge, MA: MIT Press.

Gonzáles-Bailón, S. (2013). Social science in the era of big data. *Policy and Internet, 5*(2), 147–160.

Hacking, I. (1990). *The timing of chance*. Cambridge: University Press.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Harvard: University Press.

Lazarsfeld, P. F. (1961). Notes on the history of quantification in sociology: Trends. *Sources and Problems. Isis, 52*(2), 277–333.

Lorenzet, A. (2015). Le « scene mediali » della datificazione. La controversia tecnoscientifica sulla privacy nei quotidiani italiani. *Rassegna italiana di sociologia*, LVI(3–4), 581–607.

Nature (2007, 11 October). A matter of trust. *Nature*, 449, 637–638.

Neresini, F. (2015). Quando i numeri diventano grandi: che cosa possiamo imparare dalla scienza. *Rassegna italiana di sociologia*, LVI(3–4), 405–431.

Porter, T. M. (1986). *The rise of statistical thinking. 1820–1900*. Princeton: University Press.

Supiot, A. (2015). De L'harmonie par le calcul à la gouvernance par les nombres. *Rassegna italiana di sociologia*, LVI(3–4), 455–464.

Westergaard, H. (1932). *Contributions to the history of statistics*. London: P.S. King & Sons.

Blurry Boundaries: Internet, Big-New Data, and Mixed-Method Approach

Enrica Amaturo and Gabriella Punziano

Abstract Internet as place of everyday life, as well as a vast repository of information, becomes an integral part of the society essential for understanding complex phenomena and social issues. Social researchers cannot remain anchored to traditional research practices and conceptual categories, but instead must cope with the exponential growth of new data, much of which is user-generated and freely available online. In this manner, Web-mediated research is already transforming the way in which researchers practice traditional research methods transposed on the Web. But the steady growth of interest in these new analytical frontiers has not been properly accompanied by a solid reflection on the implications that the new data available and the analysis of such data can generate. It is necessary to undertake a critical discussion of the quality of this new data and the processes involved: data collection, organization, analysis, and treatment of ethical issues. Over the next years, this domain will raise the need of to much efforts in social research including the use of new methodological approaches. Along this track, the current chapter aims to problematize future challenges for social research, with a particular emphasis on the advent of big-new data.

Keywords Big-new data • Mixed-methods approach • Internet • Online textual data • Privacy and other constraints • Data quality • Data availability • New technical frontiers

E. Amaturo (✉)
Methodology of Social Sciences, Department of Social Sciences,
University of Naples Federico II, Vico Monte della Pietà, 1, 80136 Naples, Italy
e-mail: amaturo@unina.it

G. Punziano
Urban Studies and Social Sciences, Gran Sasso Science Institute – GSSI,
Viale Francesco Crispi, 7, 67100 L’Aquila, Italy
e-mail: gabriella.punziano@gssi.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_5

1 Introduction

Nowadays, the Internet, which more and more holds a place in people's daily lives, is configured as a consideration that is essential for understanding many different phenomena and complex social issues. If research fails to take into account this vast repository for information, its potential results are of limited value. Any research that aims at making a general investigation must examine the research environment as a whole, meaning that the Internet must be an integral part. Social researchers cannot remain anchored to traditional research practices and conceptual categories, but instead must cope with the exponential growth of new data, much of which are *user-generated* and freely available online. In this manner, Web-mediated research is already transforming the way in which researchers practice traditional research methods transposed on the Web; for example, it will be considered as online ethnographic research (i.e. *netnography*) or *social network analysis* that actually can leverage a greater potential compared to relational data relying on information provided by social networks. The steady growth of interest in these new analytical frontiers has not been properly accompanied by a solid reflection on the implications that the new data available and the analysis of such data can generate. More concretely, this means that it is necessary to undertake a critical discussion of the quality of this new data and the processes involved, which include data collection, organization, analysis, and treatment of ethical issues.

First of all, there are issues of data access because of the constraints of *privacy*. Materials on the Internet, which assume the status *of personally published documents* (Amaturo and Punziano 2013), are produced and used for different purposes than analytical use (as is generally true for all secondary data). This leads to ethical and value questions over characterization of the data. Indeed, the data can be considered either as simple information that is freely available or as the result of a contextual construction that should not be investigated outside the scope in which it was produced. The ethical boundaries over use of such data are becoming increasingly blurred, as the line between public and private material also becomes blurred.

Secondly, questions arise as to the nature of the data and the ability to analyse them. The development of the Internet and smart mobile devices, with their social applications, is making obsolete the usual ways of dealing with data. This erodes the boundaries of time (synchronous/asynchronous) and space intended as boundaries between real and virtual. Consequently, new types of data have emerged, which are classified as big data (user-generated content, streaming digital data, log data, and information generated from any communicative and Internet processes).

The problems, along with that of ethics, consist in archiving, processing and analysing these new data. Over the next years, this domain will raise the need of to much efforts in social research including the use of new methodological approaches such as the *mixed-methods* approach (cf. Tashakkori and Teddlie 2010 and

Amaturo and Punziano 2016, among others) that seems to increasingly blur the boundaries between qualitative and quantitative research.

These questions represent one of the current methodological frontiers, and along this track, the current chapter aims to problematize future challenges for the approach, with a particular emphasis on the advent of big-new data. In fact, this is an approach that has stopped fighting for emergence and is now fighting for consolidation and specialization, in particular by identifying those holes where thematic areas and issues are yet missing a consolidated discussion in the current literature on mixed methods (as well as on research in general) and the challenges coinciding with the advent of big data.

2 New Contexts and New Data

This paper discusses how academic research can address real-world problems and contribute to relevant and forward-looking solutions when the context involved—and in which the data and the knowledge are generated—is the transposition on the Web of relational phenomena.

An important issue for social research concerns the diffusion of the Internet and the so-called Web 2.0 (Boccia Artieri 2015; Hesse-Biber and Griffin 2013). This is because this mechanism has changed, and at the same time innovated, the nature of the data, along with its collection methods and analytical procedures. The social research field has seen the opening of a world of opportunities and challenges with respect to the data produced and accessible from the Internet: 'A real *revolution* destined to change the ways of doing social research' (Amaturo and Aragona 2016, 26).

The fast growth of data is not a novelty. But how can we make sense of such a data explosion? This is the question with which many researchers have to fight. Web traffic and social network flows, as well as the software, sensors, and tools for monitoring these flows, have become instruments in the search for a guide to rapid and effective decision-making and learning about social processes. In contemporary society, data are constantly produced as the direct and indirect effect of bureaucratic, legislative, planning, and other activities, but also as the spontaneous accumulation of information resulting from the use of social networks as means of exchange, social relationships, and knowledge (*ibidem*).

Understanding how ideas, opinions, and behaviours are constructed, produced, and spread in large communities is a process that cannot be achieved by ignoring the Web. It is a sort of *revolution* in which 'we are really just getting under way. But the march of quantification, made possible by enormous new sources of data, will sweep through academia, business, and government. There is no area that is going to be untouched' (Lohr 2012). Nowadays, data can be considered a new class of economic asset, such as currency or gold (Chen et al. 2012), because of the knowledge power embedded therein. Data can help to combat poverty, crime, and pollution, and it can aid in understanding world changes and trajectories or assist in decision-making processes. Today, the extent of data is not only a lot greater than in the past, it also includes entirely new types of data. Data seems to be more communicative because they are generated from any digital process as well as through sensors in a completely new way. In accordance with the foregoing discussion, the *data revolution* can be defined as the process that has radically changed the routines for building, organizing, and analysing the data consolidated in scientific disciplines. This combined with developments in information technology, governance, and research techniques developments has, in turn, changed the way the data are used to produce knowledge about social phenomena (Amaturo and Aragona 2016).

According to Amaturo and Aragona (2016), in the word of new and big data, many distinctions can be made. Data can be distinguished according to the way in which they are generated: direct data (products from direct measurements; for example, recordings of surveillance cameras, aerial photographs for the monitoring of territories, etc.) are data that require human intervention to be collected and analysed; automated data, that which can be collected and produced by an automatic device (for example, data detected by a sensor related to smog, temperature, or movements in an enclosed space; or metadata about a telephone conversation, such as the recorded date, time, and duration); and volunteer data, data produced by individuals who upload or transmit data to a particular system (as in the case of images, texts, audio posted on social networks, blogs, sites, etc.) (Kitchin 2014). It is also possible to distinguish these kinds of data by the area from which the data are produced, e.g. public sector (which is generally accessible to researchers), private sector (generally used for internal investigation purposes), and users of digital infrastructures and platforms (generated by social media and thus difficult to access, organize, and analyse because they are produced for purposes other than research) (Elias 2012; Van Dijck 2009). However, one main distinction is certainly the flexibility of the data collected, for which it is possible to recognize the following categories: structured data (text and numbers arranged in relational databases or matrices, where the number of rows and columns is fixed); semi-structured data (which do not have a fixed pattern, but are organized in a flexible manner based on the available content); and unstructured data (which do not share the same format, such as data from social networks, web-sites, etc.). In this last group, it is necessary to mention the so-called *Big Corpora*, which consist of very large collections of textual data that require treatments of information extraction, cleaning and creation of the data, and—only at the end of this process—suitable organization in order that the data may be processed and analysed.

More generally, big-new data can be defined as referring to data that are *too big-large* (petabyte-scale collections of data that come from click streams, transaction histories, sensors, and elsewhere, thus entailing massive quantities of data), *too fast* (big amounts of data that also need to be processed quickly), *too complex* (to sense from social media data), and *too hard* (data that do not fit neatly into existing processing tools or that need some kind of preparation or analysis that existing tools cannot readily provide) to process for existing tools (Madden 2012). ‘The big data of this revolution is far more powerful than the analytics that were used in the past. We can measure and therefore manage more precisely than ever before. We can make better predictions and smarter decisions. We can target more-effective interventions, and can do so in areas that so far have been dominated by gut and intuition rather than by data and rigor' (McAfee et al. 2012, 62). The big data movement, like analytics before it, seeks to glean intelligence from data and translate data so as to advance knowledge in every field that can be impacted by this revolution. However, three key differences could be highlighted: the volume of the data flow, the velocity in the speed of data creation and its real-time consistency, and the variety of production sources (messages, updates, and images from social networks, readings from sensors, GPS signals from cell phones, and others) that provide enormous streams of data tied to people, activities, and locations:

> The development of the Internet in the 1970s and the subsequent large-scale adoption of the World Wide Web since the 1990s have increased business data generation and collection speeds exponentially. Recently, the Big Data era has quietly descended on many communities, from governments and e-commerce to health organizations. With an overwhelming amount of web-based, mobile, and sensor generated data arriving at a terabyte and even exabyte scale, new science, discovery, and insights can be obtained from the highly detailed, contextualized, and rich contents of relevance to any business or organization (Chen et al. 2012, 1168).

More and more high-impact application areas are involved in such radical processes, such as e-commerce and marketing intelligence, e-government and politics 2.0, science and technology, smart health and well-being, and security and public safety. Furthermore, any of these areas adopt or develop appropriate analytics and techniques to drive the intended effect of extracting knowledge from available big-new data.

Big-new data has been accelerated by advances in computing that allow us to measure phenomena in fine detail and as it happens in real time. In fact, the big-new data trend has also been fuelled thanks to the improved access to information and to the development of e-databases that archive old paper-based systems of knowledge. In this sense, data are not only becoming more available but also more understandable to computers. Words, images, video on the Web, and streams of sensor data generate a set of unstructured data that are not directly organizable in traditional databases. Consequently, computer tools for generating knowledge and insights from the flow of these unstructured data have grown fast because of rapidly advancing techniques in artificial intelligence, such as natural-language processing, pattern recognition, and machine learning. The application of these techniques has involved many fields, though it is true that some big challenges for big-new data are still far from a stable solution—such as, for example, the possibility of parsing vast quantities of data and making decisions instantaneously. Experience and intuition are progressively being replaced by decisions based on data, analysis, and empirical evidence. In this shift towards data-driven decision-making and evidence-based problem solving, the demand for greater predictive power from data has arisen in fields ranging from sporting bets to public health systems and economic development and economic forecasting.

Therefore, we are swimming in an expanding sea of data that is either too voluminous or too unstructured to be managed and analysed through traditional means. New sources, among others, are clickstream data from the Web, social media content (tweets, blogs, Facebook wall postings, etc.), and video data from retail and other settings or from video entertainment. Big-new data also encompasses everything from voice data to genomic and proteomic data from biological research and medicine. And very little of this information is formatted in the traditional rows and columns of conventional databases.

As Davenport and Bean (2012) highlight, big data offers the ability to take advantage of real-time information from sensors, radio frequency identification, and other identifying devices to understand environments at a more granular level (especially from a social sciences perspective), to create new products and services (for market competition), to treat and discover cures for threatening diseases (in life sciences), and, more generally, to respond to changes in usage patterns as they occur. Big data has become the horizon from which recovery of information is what is needed to stand apart from traditional data analysis environments. In particular, the need has now emerged in the marketplace to pay attention to data flows as opposed to stocks and to rely on data scientists and product and process developers rather than data analysts. The authors also note that big data can be useful in supporting processes (such as decision-making, prevention, and security) using real-time processed information; in involving continuous process monitoring to detect changes, trends, and trajectories; and in exploring network relationships, like suggesting friends on LinkedIn and Facebook by matching characteristics and known information. In all these applications, the data do not serve as the ‘stock’ contained in a data warehouse but rather maintain a continuous flow. A substantial change has thus occurred from the past, when data analysts performed multiple analyses to find meaning in a fixed supply of data. Given the volume and velocity of big-new data, conventional approaches to decision-making are often not appropriate in such settings. Therefore, in big-new data environments it is important to analyse, make decisions, and act both quickly and often. The new data scientist needs to understand not only analytics, but also computer science, computational physics, and biology- or network-oriented social sciences. He or she requires skills that ranging from data management to programming, mathematical and statistical skills, business acumen, and the ability to communicate effectively with decision-makers. All this goes well beyond what was necessary for data analysts in the past. A key issue for big data is the fact that the world and the data that describe it are constantly changing, and organizations that can recognize the changes and react quickly and intelligently to those changes will have the upper hand.

‘Big Data technologies […are defined…] as a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis. There are three main characteristics of Big Data: the data itself, the analytics of the data, and the presentation of the results of the analytics. Then there are the products and services that can be wrapped around one or all of these Big Data elements’ (Gantz and Reinsel 2012, 9). Clearly, one of the most relevant problems that occurs in this context is the need to bridge the gap between large-scale data processing platforms and analysis packages, but this issue involves informatics and engineering infrastructure and thus lies outside many other relevant problems. This also suggests another important aspect in processing data: the computer and mathematical models used to tame and understood data: 'These models, like metaphors in literature, are explanatory simplifications. They are useful for understanding, but they have their limits. A model might spot a correlation and draw a statistical inference that is unfair or discriminatory, based on online searches, affecting the products, bank loans and health insurance a person is offered, privacy advocates warn […] there seems to be no turning back. Data is in the driver's seat. It's there, it's useful and it's valuable, even hip' (Lohr 2012). The development of statistics and machine-learning algorithms also requires a parallel development of a consistent data management ecosystem around these algorithms, so that user can manage and evolve their data; enforce consistency properties over them; and browse, visualize, and understand their algorithms' results. Hidden behind these necessities is the fact that every system is actually collecting more data than it knows what to do with. Transforming this information into relevant and useful knowledge is a very difficult challenge.

Furthermore, 'Since the early 2000s, the Internet began to offer unique data collection and analytical research and development opportunities. […] Web intelligence, Web analytics, and the user-generated content collected through Web 2.0-based social and crowd-sourcing systems (Doan et al. 2011; O'Reilly 2005) have ushered in a new and exciting era of BIandA 2.0 research in the 2000s, centred on text and Web analytics for unstructured Web contents' (Chen et al. 2012, 1167). This means considering the data flow as a conversation that requires the integration of mature and scalable techniques in text mining (e.g. information extraction, topic identification, opinion mining, question answering), Web mining, social network analysis, and spatial-temporal analysis. The developments and advancements in these areas have had an impact on both industry and the academic sector, leading to a progressive convergence between the two.

In this revolution, another element has become noteworthy: 'At the same time, the steadily declining costs of all the elements of computing—storage, memory, processing, bandwidth, and so on—mean that previously expensive data-intensive approaches are quickly becoming economical' (McAfee et al. 2012, 63). A new frontier has been opened; large amounts of information exist on virtually any topic of interest, and each of us is now a walking data generator. This statement illustrates the power of big data to inform more accurate predictions, better decisions, and more precise interventions, and to enable these things on a seemingly limitless scale (ibidem). Big-new data's power does not erase the need for vision or human insight nor does it negate the role of intuition and experience in the decision-making process.

There is, also, excitement about big-new data technologies, automatic tagging algorithms, real-time analytics, social media data mining, and the myriad of new storage technologies. In addition, big-new data are already transforming the study of how social networks function and what kinds of information they can generate accidentally. The development of sentiment analysis, for example, has greatly enriched the fields of advertising and marketing, politics, and public opinion management, not only in matching people to interest or people to people, but also in creating huge online data sets of collective behaviours and opinions, from which it is possible to retrieve the necessary information to develop knowledge systems on the changing society.

Big-new data environments must make sense of new data. As big-new data evolves, the architecture will develop into an information ecosystem, a network of internal and external services continuously sharing information, optimizing decisions, communicating results, and generating new insights. Furthermore, we are used to thinking of big-new data as always telling us the truth, but this is actually far from reality.

Some problems occur in data acquisition (determining whether all the data are important, or how to filter out the right information) and in extraction of information (pulling out the required information from its underlying sources and expressing it in a structured form suitable for analysis), and there also remains an important question of a suitable database design that permits differences in data structure and semantics to be expressed. As was mentioned earlier, one problem with current big-new data analysis is the lack of coordination between database systems and analytics packages that perform various forms of data processing, such as data mining and statistical analyses. Moreover, the ability to analyse big-new data is of limited value if users cannot understand the analysis: 'There is a multistep pipeline required to extract value from data. Heterogeneity, incompleteness, scale, timeliness, privacy, and process complexity give rise to challenges at all phases of the pipeline. Furthermore, this pipeline is not a simple linear flow—rather there are frequent loops back as downstream steps suggest changes to upstream steps. There is more than enough here that we in the database research community can work on' (Labrinidis and Jagadish 2012).

Inside this greater context of concerns, one relevant question is linked to privacy. Amidst the enormous increase in digital data and transactional information, concerns about information privacy have emerged globally: 'Privacy issues are further exacerbated now that the World Wide Web makes it easy for the new data to be automatically collected and added to databases' (Agrawal and Srikant 2000, 439). These concerns come from the previous Internet era, in which it was an established fact that people disguised some basic characters behind false identities because of misunderstandings or accentuated concerns about privacy. The fact that such people had then, in this context, acted upon information related to their true selves showed that they were not equally protective of every field in their data records. Such disguises have become less important for people, organizations, and everything else occurring on the Internet today. The big-new data stream and flows are further and further from concerns arising from the possibility of veiled identities, actions, and patterns of usage.

What we can say is that a new context for information, the so-called digital universe, has generated a new tangible and digital geography (Gantz and Reinsel 2012), which is conceivable as a vast, barely charted place full of promise and danger. A majority of information in the digital universe is created or consumed by consumers, but the amount of information individuals create themselves (writing documents, taking pictures, uploading) is far less than the amount of information being created about them in the digital universe (ibidem). Because of that, issues strongly emerge concerning copyright, privacy, and compliance with regulations even when the data zipping through networks and server farms are created and consumed by consumers. In a report titled ‘The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in Far East’, much attention is paid to the data producers and, as many producers are essentially consumers on the Internet, much of the data in the digital universe is unprotected. The question arises, then, whether some type of protection is required for such data (to protect privacy, adhere to regulations, or prevent digital snooping or theft): ‘Therefore, like our own physical universe, the digital universe is rapidly expanding and incredibly diverse, with vast regions that are unexplored and some that are, frankly, scary’ (Gantz and Reinsel 2012, 4). The need for information security is growing fast because of the rise in mobility and participation in social networks, the increasing willingness to share more and more data, and the new technologies that capture more data about data (the so-called metadata). For data that may have some level of sensitivity, there are several levels of security definable as privacy only (such as an email address or a stream of downloads), compliance-driven (such as emails that are contentious or subject to storage rules), custodial (such as account information potentially subject to identity theft), confidential (information protected from the origin such as personal memos, transactions, and so on), and lockdown (information requiring the highest security, such as financial transactions, personnel files and documents, medical records, etc.). In this sense, the issue may be more sociological or organizational than technological (ibidem). Also, if our digital universe is bigger and will be bigger, more valuable, and more volatile than ever, one solution contemplates the harmonization of law and regulations governing information security around the globe. As digital data transcend conventional boundaries, a global knowledge of information security may be the difference between approval and denial of a data request.

3 The Study of the Online Textual Data[1]

3.1 What Is New in Approaches and Analytics?

Given the huge amounts of data available and the different nature that different sources of data can assume, in the following section, we want to develop a more narrow discourse concerning big-new data produced by the interaction and Internet transposition of relational phenomena, although it is true that the impact of data abundance coming from the Web extends well beyond social processes and also involves the economic, business, political, and many other fields. This combination of methods, approaches, and scientific disciplines has to be seen as an opportunity,

[1] The reflections in this section recall the articulate discussion in Punziano (2016, 149–157).

because the new demand for knowledge from the Web, and from society itself, is becoming increasingly data-intensive. In this scenario, the real concern is not only the possibility to apply a computer-automated analysis, but the ability to govern it.

Moreover, in the transition from analogue to digital data, the study of content carried on the digital network in addition to the abundance of available data also established a very happy marriage with the development of automated analysis in one field of study, that of *content analysis*, which had often been accused of being strongly subjective and particularly expensive in terms of time and resources. Much progress has been made in terms of computer programming and hardware improvement, but the most essential improvements have been in textual statistics, machine learning, and on advanced and intuitive methods of visualization that allow for analysis without having to resort to a priori assumptions that are thus very iterative. This makes such systems both fundamental and extremely valid from the point of view of exploratory analysis, while presenting non-trivial difficulties from a computational point of view. This moves the subjective limitations of content analysis towards objectification, even if significant barriers are still found. However, the extensive development of available software has prompted an enormous acceleration in the use of documents as sources for social research, thus triggering the perverse mechanism of an exponential growth of additional software and applications. This direction of development has meant that the trend towards automation of analytical systems has led to a continuous approaching of quantification targets and the extension of qualitative studies, as well as to a progressive increase in-depth pretention in quantitative studies. This ensures that both approaches have seen advantages in terms of the ability to handle rapidly and easily a large amount of data, by giving to the entire analytical process a greater scientific rigour, database available for inspection, and reproducibility of the analytical procedures. However, both approaches have equal value in the cognitive scope with respect to a given phenomenon, bringing the results closer from different angles. As a result of the complexity of the phenomena under investigation in social research, and particularly in communication and politics studies, it is necessary to combine these two approaches so that the weakness of one can be overcome by the strengths of the other (Denzin 1978). This is still a difficult task, as the results obtained, in order for them to be read together, require the definition of a common vocabulary and an approach of two totally different ways of thinking, or rather must be thought of as opposites.

Therefore, if quantitative methods are used to obtain reliable models, qualitative methods are used to capture the essence of a phenomenon. The *blurring of boundaries* and the integration into routine analytical procedures, however, leads to the emergence of a third analytical approach, which aspires to establish itself as an autonomous and independent approach in social research: the *mixed-methods* approach. This approach integrates the qualitative and quantitative visions in three dimensions—epistemological, ontological, and methodological—leading to an integrated but not necessarily convergent approach. It does not use the different methods in the form of triangulating the results in order to assess their consistency, but instead uses the methods as part of an investigation toolbox that can broaden the vision of certain phenomena and may allow for the emergence of significant and independent results (Crassewel 2003; Taschakkori and Teddly 2010). Basically, the mixed-method approach is an interesting approach devoted to integrating methods, data, and researchers. It takes the form of an approach to theoretical and practical knowledge that attempts to consider multiple points of views and perspectives, including both qualitative and the quantitative ones (Johnson et al. 2007). It is not, however, merely a summation of approaches and analytical methods, but rather an integrated and integral way of looking at the investigated reality (Bryman 2004) which allows researchers to overcome the limitations of individual approaches in order to achieve a complex model, along with a container and amplifier, of the methods individually used (Denzin 1978). Hesse-Biber and Johnson (2013), describe the mixed strategy as a means of ‘coming at things differently’ (Hesse-Biber and Johnson 2013, p. 103). The traditional forms of data collection, using only one method at a time, may not be adequate to address complex questions that sometimes require a variety of qualitative and quantitative lenses to achieve the desired optimal results from the study. The purpose of this approach is to improve the width, depth, and complexity of the produced knowledge (Daigneaut and Jacob 2014). In addition, it shows itself as dynamic, flexible, and inclusive, allowing in this way the deepening of complex and strongly changeable phenomena as those that take shape on the Internet.

The study of social and communication phenomena through *online textual data* is a new frontier and a new challenge in understanding society in its transposition on the Internet. But in order to make this study possible, it is necessary to consciously consider and identify the limits and new prospects for the use of online data in social research.

In particular, the use of online social text data becomes ever more essential for understanding and analysing new forms of participation in the flow of communication on the Internet. However, it is also important for the researcher to investigate the quality of online data, the ways in which these data are produced, and the possibilities that they offer to answer different research questions. This requires an acute reflection on the role of social research techniques ‘on’ and ‘with’ the Internet, that is, the so-called *Web Content Analysis* (Lilleker and Jackson 2011).

WCA, as an analytical perspective, has shown potential for development at a time of crisis between the underlying approaches to content analysis. Jointly to the emergence of the Web, which has dramatically expanded the availability of data, serious problems have arisen in relation to cognitive questions and this abundance of content and its conveyed meanings, as well as the selection, veracity, and traceability of information sources, and the use and consumption of these generated materials. The approach to statistical techniques or other quantitative analyses in this area has made it possible to dominate and systematize the availability of analytical objects by creating specific characterizations, compared to the analysis of the content published on the Web, or the *WCA*. That prospect has particular features that show points of similarity with the general technique from which it results, content analysis, but it reflects the advantages and disadvantages of the Web as an object and as a place of analysis. After all, the growing attention to the Internet and its content lies in the development from a one-to-many transmission model to many-to-many—in this way, changing the rules of the game in the classic communication phenomena. In its evolutions, WCA uncovers this ambiguity and leads to a distinction involving two particular families of techniques: the traditional application of the techniques of content analysis to the Web interpreted as *strictly* in this way (McMillan 2001), and the analysis of Web content through specific techniques (Mitra and Cohen 1999; Wakeford 2000) that allow for systematic identification and study of linkage models, of message characteristics with interactive content, etc. It was specifically for the analysis of messages produced by interaction that an intense strand of computerized analysis studies of speech was developed, as described by Herring (2004), while for understanding the hypertext nature of websites was born the interconnections models (formed by hyperlinks), which dealt with the classical methods of social network analysis (SNA).

In addition, a macro-division can be carried out between: (i) *ethnography of the Web*, on the qualitative side, focusing the analysis on Web content essentially as sociocultural texts and describing and exploring them from a qualitative point of view to discover the basic meaning; and (ii) *analysis of content and functionality of the Web*, which focuses on the web page by viewing sites such as units of analysis, and developing a coding system that is applied to them in order to provide quantitative measurements of the content features, functionality, usability, and design. There is also a mixed-methods form of content analysis on the Web, whereby advocates Gibson (2010) and Lilleker and Jackson (2011) have developed mixed integrated qualitative and quantitative methods for the study of electoral campaigns on the Web, showing that, with the development of Web 2.0, it is impossible to think of static approaches in extension (quantitative) or in-depth (qualitative), but instead there is a need to look at nonparametric dynamic metrics that model the analysis at the speed of change inherent in the Web itself. Also important is the development of interactive methods based on control of a vast database and large samples, finalized by comparative analysis and with the emergence of standardized measures and automated features for the acquisition and analysis of data from social media.

3.2 Concerns and Boundaries

The concerns raised by the innovative reach of Web content and the methods to analyse such content are a necessary part of the process in the evolution of a research paradigm and in the need for *scientific rigour* of new approaches and new techniques associated with these developments.

Concerning the nature of the data on the Internet, the data collected from social networks are often ambiguous, problematic, and of uncertain authors, with insufficient analysis of the content carried on the Internet. Classical analysis techniques are not very suitable to enter a scene with this level of complexity, because these techniques lead the researcher to focus on the data itself rather than on the potential of the container and the opportunities that it offers in terms of resonance. The limits of the classical analysis techniques reside precisely in their inability to reflect the ambiguities and potential scenarios opened by the Internet. The contribution of classical analysis techniques remains valid so long as the data to be analysed are highly repetitive and in a 'manageable' dataset. Such techniques are therefore not able to explain the potential of interactivity inherent to the social Web. These are techniques that generally do not work on relational matrices or interconnections between the data produced, but they allows for analysing the overall data extrapolated from the container (the Internet). This leads to the need to fragment a production scenario, which in itself presents a strong complexity and is difficult to read whether decomposed and simplified through the use of a few variables in a descriptive sense. The scenario or container, then, and the different production practices—use and generation of information for social users, as the producers, consumers, and reusers of the information transmitted through the investigated channel—are elements that the classic techniques continue to sacrifice because they are unfit to grasp the particularity of these items. New horizons have unfolded, however, thanks to the introduction of technical specifications for analysis on the Internet born from the union of text mining developments, applications of neuronal networks, big data processing through machine-learning systems, and in general from the emergence of data science as a possible applicative approach for statistical analysis in social research. Fruitful applications that demonstrate the potential to develop specific techniques for the Internet have been produced in the field of content analysis. Consider, for example, applications of *social network text analysis* (Ehrlich et al. 2007) which, unlike the analysis of classically understood content, is directed to the reconstruction of networks of relationships in which only certain concepts or areas of meaning are discussed and disclosed. Another analytic boundary is the analysis of short texts on social channels, which could lead to overcoming the lack of depth of content analysed by classifying them with respect to the polarity of associated opinions and expressing this as *sentiment analysis* (Pang and Lee 2008). This again responds to a new cognitive objective that goes beyond the mere content, or the structure of relationships that drive the content, beyond the scenario, and seeks to retrieve user practices and the interactive practice inherent in the social network. A further reflection can be highlighted. What is happening in today's society is that often specific functions end up being attributed to each platform that is used; this goes from identity spheres to showcases, from accurate reproductions of social worlds within the social network to completely virtual worlds with ever more realistic games. It is through the strong connotative use of Web channels that messages and, more generally, content are conveyed, and thus in the kind of social online data that derive from these. This is the whole crux of the matter regarding the methodological implications in the use of these specific data and in the necessary associated analysis techniques. The limits and potential of the proposed analyses and of the emerging analytical boundaries are still open for debate. These new techniques offer the possibility of jointly analysing both the content—as well as the way in which it is used and reused—and the container in which it all takes place.

Additionally, with the advent of the Internet, additional ethical issues are generated that can also become quite thorny, for *mixed-methods* as well as for social research in general. Such is the case with the use for research purposes of *personal documents published on the Web* (Amaturo and Punziano 2013, p. 88) without those who produce them being aware of the fact that such documents can become the object of a potential study. And again, the development of Web 2.0 and smart mobile devices, with their increasing number of social applications, may soon make obsolete methods to process data that confuse the *boundaries of time* (synchronous/asynchronous) and *of space* intended as *boundaries between real and virtual*. *Ethical boundaries* become increasingly murky as the *boundary between public and private* blurs. In addition to these ethical concerns, the greatest emerging problems, given the huge amount of product data in the network, are the storage, processing, and analysis of these big-new data, and these problems in the coming years are likely to drive mixed-methods researchers to rethink, innovate, and produce new paradigmatic perspectives, as well as new structures and research designs:

> The exponential growth of "big data", arising from newly emergent user-generated and streaming digital data from networking sites such as Twitter and Facebook, will place pressures on MM researchers to transform traditional modes of collecting and analyzing data generated from these sites. Big data "is [sic] too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this [sic] data, you must choose an alternative way to process it [sic]" (Dumbill 2013, p. 1). How will MM researchers incorporate big data and ''big analytics'' (i.e., large-scale algorithms developed to understand the meanings contained within individual social networking outputs)? In the coming years, big data methods and analytics may also drive and challenge MM researchers to rethink and innovate and produce new paradigmatic perspectives and research designs and structures. In turn, MM perspectives and praxis can provide models for interpreting and deriving critical insights that that may give a more complex understanding of big data that can bring a set of new questions and understanding to the trending data currently extracted from user-generated social networking sites (Hesse-Biber and Johnson 2013, p. 107).

4 Open Questions[2]

Regarding what has been discussed in the previous sections, and taking up the arguments of Hesse-Biber and Griffin (2013), one issue still under exploration for mixed-methods appears to be the use of technology mediated by the Internet for data collection. Advantages, disadvantages, and ethical issues are all factors involved in this process, and it is impossible to disregard these in discussing how

[2] The reflections in this section recall the complex debate exposed in Punziano (2016, 157–162) and its developments in Amaturo and Aragona (2016, 25–50).

the Web has changed the possibilities of movement for the researcher and opened scenarios of knowledge that were roughly unthinkable until now.

The Web 2.0 is a multifaceted scenario, but it is essential to understanding many phenomena and very complex social issues. Apart from noting that to ignore this boundless container of information limits the potential value of research, advocates of mixed-methods research have for decades argued that adopting a single approach, either qualitative or quantitative, confines the possibilities of knowledge. Even the reality to be investigated should be examined as a whole, and every day, the Internet becomes more and more the space where this information ends up.

Just to cite a few figures, consider that, according to the Internet World Statistics (2015), about 30% of the world's population are Web users and, in Europe and America, about 80% of households are connected to the Internet, with access rates ranging from weekly to daily. It is not hard to guess how pervasive the growth of user-generated data has become, and it is therefore ever more limiting for the research community to remain anchored to the traditional practices and concepts. Thus, Internet-mediated research is already transforming the way in which researchers practice traditional research methods transposed on the Internet; for example, important illustrations include the previously mentioned ethnographic online research (Andrews et al. 2003; Denissen et al. 2010; Dicks and Mason 2008; James and Busher 2009; Robinson and Schulz 2011), which nowadays has the more specific connotation of *netnography* (Kozinets 2010; Addeo and Esposito 2015), and *social network analysis* (Wasserman and Faust 1994; Scott 2012), which currently can make the most of its potential compared to relational data by relying on information provided by social networks (Steinfield et al. 2008).

The steady growth of interest in research mediated by the Web, however, has not been adequately complemented by solid reflection and careful consideration of the extent to which this can actually facilitate the practice of mixed methods, as well as of the other individual approaches. Experiments, comparisons, and demonstrations have proliferated, serving to set up a complex mosaic that still needs to find the right linkage. Hewson (2003, 2007, 2008), reflecting on the contribution provided by Web-mediated research using the mixed-method approach, observes that finding a basin of data collection directly from the Internet allows the researcher to have a more dynamic element for validating research results obtained offline, or at least provides complementary elements to fill the gaps from offline collection when the investigated phenomena also have a significant component online (Davis et al. 2004). Online collection may also assist in defining the most representative samples, which can improve the ability of researchers to generalize the overall results of their studies (Hewson 2003).

In contrast to the exciting possibility that this online scene offers to researchers, there are also quite a few brakes. As is known, there is a wide selection of classical literature on the influence of the medium through which a message is sent; consider the contributions of McLuhan (1964) and Meyrowtz (1994) and of the famous statement: 'the medium is the message'. With the advent of the Internet, there are not insubstantial concerns about the influence exerted by this new medium with respect to information and produced communication, as the technology ends up structuring an environment with its own rules and thus a process takes place that only in appearance looks like what is happening offline. In seeking to enter into these dynamics, know them, and force them into a given system of interpretation, the researcher can no longer ignore the characteristics and rules of this environment.

Although many disciplinary fields have adopted the Internet-mediated research, there remain many obstacles and concerns for its development in the social sciences. First of all, there are the *questions of access to data* for the *constraints of privacy*, because much of the material produced on the Internet is produced and enjoyed for purposes other than analysis. These are *ethical questions* about the value assigned to the data, in particular whether it can be considered as simple information that is freely available or whether it is the result of a contextual construction that should not be investigated outside the scope in which it was produced. In other words, the Internet has obscured the *ethical boundaries* between what is public and what is private. In addition, many prejudices remain about the veracity and plausibility of data circulating on the Internet, the inability to verify the source of the message, and the reliability of the content. Numerous problems also arise regarding the *issues of representativeness of the achieved sample*, given the changing nature of *cyberspace*, where constancy is as unrealistic as the pretention to photograph a moment, if immediately before and immediately after that moment, the image can change with impressive speed. This has strong implications for the possibility of generalization to a population that is also changing, but not as quickly (Andrews et al. 2003). And again, data produced in very different ways requires a researcher to leave his or her disciplinary zone of comfort, consisting of analytical and technical certainties, in order to develop new skills with respect to both the collection and the analysis of new data. And if this particular point scares disciplinary and methodological purists, it will find mixed-methods researchers more ready to incorporate these changes and new challenges, given the flexibility that the approach in question requires.

Despite the limits and issues already presented for online research, there are many strengths that push towards the value of further development with a goal of overcoming these limits. Among the advantages is the *overcoming of spatial barriers* in data collection, which allows researchers to work on large samples with extremely limited costs; the online environment permits immediate accessibility that classical instruments cannot grant to researchers. Nevertheless, researchers should ignore the need for reflection on the possibilities of compliance and the return of the information, on the delimitations of the real space in which the respondents are located, and especially on who are the real users of the Internet, the entire population or only the part that is more educated, younger, and familiarized with new technologies? Another strength is the ability to leverage both synchronic collection methods, i.e. collecting *streaming data* (conversations, comments, posts, audio-visual material, and anything else that is freely available on social platforms, or even retrieving something closer to traditional data, such as articles on platforms of information) and asynchronous collection methods (email inquiries, questions to Delphi, or online surveys). Whatever the method chosen, though, one effect is the progressive distancing of the researcher from the situation of a face-to-face collection method, which decreases the influence that the researcher's presence may have on the data that he or she intends to collect. However, the absence of non-verbal signals in online interactions can become a new obstacle that counters the implicit advantage just emphasized. The impact of the Internet is, therefore, extremely controversial, and whenever there is an advantage to its use in data collection and research, in general, it tends to immediately arouse new doubts and controversy. It is for this reason that Hesse-Biber and Griffin (2013) have questioned whether researchers are really aware of the potential impacts that the new technological mode of mediation can have on every aspect of the research process and what new questions such awareness can bring. The authors, finally, identify points of strength that are more or less common in studies developed in the field of mixed-methods research making use of Internet mediation, though this does not go so far from the advantages that can be derived from all the social research implemented in the Internet. Summarizing what has already been discussed, it is essential to take into account:

1. the possibility of reaching populations that are otherwise hard to contact through the identification of groups of interest, and the possibility of achieving a wide range of cases that are territorially dispersed;
2. the increase of sensitivity and fidelity of data, especially if what is under investigation involves unknown phenomena, and the ability to identify networks that have influence in the formation and definition of opinions;
3. the gains in terms of time (such as platforms for online surveys with questionnaires that allow responses to be recorded directly in a data matrix, operations that reduce time for insertion, eliminate potential human error in transcription, and thus automatically enhance the validity of the process) and cost deduction (e.g. cost reductions for travel or office materials such as paper, pens); reducing the time factor in mixed-method sequential or concurrent studies helps to reduce collection times for different data and thus may potentially restrict one of the basic problems of this approach, namely the opportunity to collect both sets of results, discuss them jointly, and publish them in studies that are not fragmented by the availability of first one set of data and then the other; and
4. the potential increase in participation rates and responsiveness thanks in part to the simplicity of the built tools (e.g. online questionnaires), but also because it is possible to answer extremely sensitive questions while masked behind a monitor (increases in terms of the perception of anonymity may consequently decrease the social desirability bias of responses), in the safety of the respondent's own home, and with the ability to use as much time as necessary to provide complete answers to complex questions.

This last point exactly can be a double-edged sword. In fact, the advantages listed are surely matched by the decline in terms of empathy. An online interview does not have the same advantages, in terms of involvement, of an interview conducted in person, and it can create a sense of disinterest in the respondent if he or she ends up attributing little weight, relevance, and scientific value to an analysis conducted in this way; such a respondent may also fail to pay due attention to the answers given because he or she considers them of little use in achieving their desired purpose. In this sense, the researcher, due to the lack of interpersonal dynamics, will have a more difficult task in establishing an online relationship rather than an in person one. But more than an increase in terms of process validity, validation of results, and verification of authenticity of the obtained results, the use of online research becomes instrumental in improving the overall understanding of the research problem. Before designing an online research effort, a researcher must have very clear objectives for the research, an understanding of the effects of Internet mediation, and knowledge of how all this can improve the overall research process by limiting the *side effects of online transposition*. One of the biggest problems continues to be that of the under-representation on the Internet of some specific population types (e.g. the homeless, the elderly, and the less-few educated people), so to the increase in terms of possible contact in a large space with reduced time and costs, it goes to take into account the need, sometimes essential, to support online procedures with the classic offline procedures, whose combined use becomes very useful at this stage of advancement of the process of computerization of society. The Internet permit to the social scientists to open new scenarios to provide meaning to complex social environments inside and outside the Internet itself, leading to new methodological challenges and new stimuli for the imagination of the researcher through new research directions.

As seen, the online research is not devoid by drawbacks. For example, Hine (2008) questions the real nature of collected data online and shows decided doubts about the possibility of attributing to an Internet study the characteristics of a study in-depth, as the Web offers a dynamic and fleeting environment as its intrinsic connotation. Among the drawbacks, synthesizing, there are:

1. the lack of generalizability, because Internet users tend to be still mainly composed of more educated people, young, and affluent enough to afford access and tools;
2. the digital divide and technological development, for whom the researcher will have to keep in mind the technological developments and the extent to which they reach the population as a whole before being certain of potentially accessible from the online research coverage level, also bearing in mind that these parameters can vary from country to country and that the exclusion of a significant part of the population by social studies in the Internet can become a great threat to the randomness and representativeness of the investigated samples;
3. the self-selection of respondents and often uncontrollable multiple compilations that ultimately affect the generalizability of the results of the study;
4. the ethical issues concerning the establishment of identity on the Internet and the basic sociodemographic characteristics essential to the performance of any social study such as age, sex, race, or other personal demographic characteristics; this information depends entirely on the honesty of the respondents;

5. the lack of non-verbal cues and the difficulty in interpreting the silence and temporal fluctuations in responses that, unlike it happens in face-to-face interaction, the researcher cannot base his knowledge on the non-verbal support and the interpersonal connections that help establish rapport such as tone of voice, body language, gestures. This, in addition to affecting the wealth of qualitative data in the online research, also requires the researcher to learn how to assess other elements that can give colour to the collected data, such as the emoticons, the use of punctuation, the online jargon, that can add emotion to online interview. Interesting is the interpretation of silence, so Madge and O'Connor (2002) provide a number of alternatives to the intended meanings of silence as a delay in the digitization of the thoughts: the interviewee may need time to process; he forgot to press the sending; he has no clear ideas or did not understand the question that is asked to him; or simply, he is just slower and less comfortable to type rather than verbally expressing his thoughts.

All these problematic issues lead to the need that researchers are properly trained with respect to the research mediated by the Internet in order to be able to conduct not only research, but also to know how to interpret the results. Researchers should be aware of the possibilities of research on the Internet, but also of the available options to them as a plausible alternative research paths. Along with the steady progress of opportunities in terms of research offers by the Internet, innovative efforts have to be implemented in research, especially in the mixed one that permit to provide revolutionary answers through the integration of multiple paths. And this without forgetting that many of the traditional ethical concerns remain unchanged both if the research is conducted online than if it takes place offline. Also, the boundary between what is public and what is private becomes increasingly blurred and the possibility of access to research arena without reveal their intentions, even if it becomes a very easy choice for access to Web-community and social network, it should be equally considered and adequately justified, not only on the ethical side, but also on the methodological one.

References

Addeo, F., & Esposito, M. (2015). *Informal learning and ldentity formation: A case study of an Italian virtual community*.

Agrawal, R., & Srikant, R. (2000, May). Privacy-preserving data mining. *ACM Sigmod Record*, *29* (2), 439–450.

Amaturo, E., & Aragona, B. (2016). *La "rivoluzione" dei nuovi dati: quale metodo per il futuro, quale futuro per il metodo?*. In F. Corbisiero and E. Ruspini (Eds.), *Sociologia del futuro. Studiare la società del ventunesimo secolo* (pp. 25–50). Milano: CEDAM.

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

Amaturo, E., & Punziano, G. (2013). *Content Analysis: tra comunicazione e politica*. Milano: Ledizioni.

Andrews, D., Nonnecke, B., & Preece, J. (2003). Electronic survey methodology: A case study in reaching hard-to-involve Internet users. *International Journal of Human-Computer Interaction, 16,* 185–210.

Boccia Artieri, G. (2015). *Gli effetti sociali del web*. Franco Angeli: Forme della comunicazione e metodologie della ricerca on-line. Milano.

Bryman, A. (2004). *Social research methods* (2nd ed.). Oxford: Oxford University Press.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence And Analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188.

Cresswell, J. W. (2003). *Research design: Qualitative, quantitative and mixed method approaches*. London: Sage.

Daigneault, P. M., & Jacob, S. (2014). Unexpected but most welcome mixed methods for the validation and revision of the participatory evaluation measurement instrument. *Journal of Mixed Methods Research, 8*(1), 6–24.

Davenport, T. H., Barth, P., & Bean, R. (2012). How big data is different. *MIT Sloan Management Review, 54*(1), 43.

Davis, M., Bolding, G., Hart, G., Sherr, L., & Elford, J. (2004). Reflecting on the experience of interviewing online: Perspectives from the Internet and HIV study in London. *AIDS Care, 16,* 944–952.

Denissen, J. J. A., Neumann, L., & van Zalk, M. (2010). How the Internet is changing the implementation of traditional research methods, people's daily lives, and the way in which developmental scientists conduct research. *International Journal of Behavioral Development, 34,* 564–575.

Denzin, N. K. (1978). *The research act*. New York: McGraw-Hill.

Dicks, B., & Mason, B. (2008). Hypermedia methods for qualitative research. In S. N. Hesse-Biber & P. Leavy (Eds.), *Handbook of emergent methods* (pp. 571–600). New York, NY: Guilford Press.

Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Communications of the ACM, 54*(4), 86–96.

Dumbill, E. (2013). Making sense of big data. *Big Data, 1*(1), 1–2.

Ehrlich, K., Lin, C. Y., & Griffiths-Fisher, V. (2007). Searching for experts in the enterprise: combining text and social network analysis. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work* (pp. 117–126).

Elias, P. (2012). *Big data and the social sciences: a perspective from the ESRC*, presentation at the conference Shaping society.

Gantz, J., & Reinsel, D. (2012). *The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east* (pp. 1–16). IDC iView: IDC Analyze the future.

Gibson, R. (2010). *Web content analysis*. Manchester: University of Manchester Press.

Herring, S. C. (2004). *Computer-mediated discourse analysis: An approach to researching online behavior*. Cambridge: Cambridge University Press.

Hesse-Biber, S. N., & Griffin, A. J. (2013). Internet-mediated technologies and mixed methods research problems and prospects. *Journal of Mixed Methods Research, 7*(1), 43–61.

Hesse-Biber, S. N., & Johnson, R. B. (2013). Coming at things differently future directions of possible engagement with mixed methods research. *Journal of Mixed Methods Research, 7*(2), 103–109.

Hewson, C. (2003). Conducting research on the Internet. *The Psychologist, 16,* 290–293.

Hewson, C. (2007). Web-MCQ: A set of methods and freely available open source code for administering online multiple choice question assessments. *Behavior Research Methods*, *39*, 471–481.

Hewson, C. (2008). *Internet-mediated research as an emergent method and its potential role in facilitating mixed methods research*. In S. N. Hesse-Biber and P. Leavy (Eds.). *Handbook of emergent methods* (pp. 543–570). New York, NY: Guilford Press.

Hine, C. (2008). Internet research as emergent practice. In S. N. Hesse-Biber & P. Leavy (Eds.), *Handbook of emergent methods* (pp. 525–541). New York, NY: Guilford Press.

Internet World Statistics (2015). *Internet users and population stats for the Americas*.

James, N., & Busher, H. (2009). *Online interviewing*. Thousand Oaks, CA: Sage.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Towards a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112–133.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. London: Sage.

Kozinets, R. V. (2010). *Netnography: Doing ethnographic research online*. London: Sage Publications.

Labrinidis, A., & Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment, 5*(12), 2032–2033.

Lilleker, D., & Jackson, N. (2011). Campaigning. Elections and the Internet: US, UK, Germany and France. London: Routledge.

Lohr, S. (2012). The age of big data. *New York Times*, *11*.

Madden, S. (2012). From databases to big data. *IEEE Internet Computing, 3,* 4–6.

Madge, C., & O'Connor, H. (2002). Online with e-mums: Exploring the Internet as a medium for research. *Area, 34,* 92–102.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data. The management revolution. *Harvard Bus Rev, 90*(10), 61–67.

McLuhan, M. (1964, tr.it. 1994). *Understanding media: The extensions of man*. Cambridge: MIT Press.

McMillan, J. H. (2001). *Essential assessment concepts for teachers and administrators*. Thousand Oaks, CA: Corwin Publishing Company.

Meyrowitz, J. (1994). Medium theory. In D. Crowley & D. Mitchell (Eds.), *Communication theory today* (pp. 50–77). Cambridge: Polity Press.

Mitra, A., & Cohen, E. (1999). Analyzing the web: Directions and challenges. In S. Jones (a cura di). Doing internet research. London: Sage. 179-202-220.

O'Reilly, T. (2005). *What is Web 2.0? Design patterns and business models for the next generation of software*.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval, 2*(1–2), 1–135.

Punziano, G. (2016). Il futuro dell'approccio: vantaggi, limiti e nuove prospettive. In E. Amaturo & G. Punziano (Eds.), *I Mixed Methods nella ricerca sociale*. Roma: Carocci Editore.

Robinson, L., & Schulz, J. (2011). New fieldsites, new methods: New ethnographic opportunities. In S. N. Hesse-Biber (Ed.), *The handbook of emergent technologies in social research* (pp. 180–198). New York, NY: Oxford University Press.

Scott, J. (2012). *Social network analysis*. London: Sage.

Steinfield, C., Ellison, N. B., & Lampe, C. (2008). Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of Applied Developmental Psychology, 29*(6), 434–445.

Tashakkori, A., & Teddlie, C. (2010). *Handbook of mixed methods in social and behavioral research*. London: Sage.

Van Dijck, J. (2009). Users like you? Theorizing agency in user-generated content. *Media, Culture and Society, 31*(1), 41.

Wakeford, N. (2000). *New media, new methodologies: Studying the web*. In D. Gauntlett (a cura di). *Web.studies: Rewiring media studies for the digital age* (pp. 31–42). London: Arnold.

Wasserman, S., & Faust, K. (1994). *Social network analysis. Methods and applications*. Cambridge: Cambridge University Press.

Social Media and the Challenge of Big Data/Deep Data Approach

Giovanni Boccia Artieri

Abstract As sociologists and media scholars, we have been facing an explosion of data regarding every aspect of our everyday life. We observe the human behavior through this large amount of data that we distribute online as contents and conversations (on Facebook, Twitter, etc.). Content created by users online (UGC) is a new kind of research topic in the field of social sciences, and it gives us particularly promising data. The digital nature of the information, on the one hand, and their enormous quantity on the other hand, change the phase of recovery and structure of the research data within social media. In other words Big Data have been changing the way we are thinking and conducting the research: we are able to analyze in depth a wide range of data as never before, facing a scenario rich both of opportunities and critical points to not understate . To observe this reality means to take into consideration a complex scenario, which is characterized by the contingency of these processes, the affordances of different technological platforms, the networks created by friends/followers, the topical networks generated, for example, by #hashtags, the structure of conversations, and the meaning of generated contents. At the same time, we need to think the relationships between Big Data and Deep Data and consider in which ways the data quality present in the quantitative aggregation (that is, the Deep Data present in the Big Data) is structurally different from the one graspable, for example, through ethnographic approach and in-depth interview techniques.

Keywords Computational social science • Social network sites • User-generated content • Big Data • Methodology

G. Boccia Artieri (✉)
University of Urbino Carlo Bo, Urbino, Italy
e-mail: giovanni.bocciaartieri@uniurb.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_6

1 The Field of Observation: The Nature of the Data and the Perspective of Collection

As sociologists and media scholars, today we have been facing an explosion of data regarding every aspect of our everyday life. We observe the human behavior through this large amount of data that people distribute (spread) online as contents and conversations (on Facebook, Twitter, Instagram, etc.). Content created by users online (UGC) is a new kind of research topic in the field of social sciences, and it gives us particularly promising data.

We are facing with a mutation in research that we can observe reflecting on how the nature of the data changes and how its collection is modified.

From a theoretical point of view, content produced online benefits from a series of ownership that defines the possibilities within the communication.

Persistence, spreadability, visibility, and searchability (Boyd 2008) of UGC allow us to observe and analyze the contents and structure of online conversations in ways which were previously impossible.

By analyzing the intersection of these characteristics, it is possible to understand the level of radical breakdown that they entail compared to the previous forms of contents produced, for example, by audience.

The persistence of the digital content on SNSs modifies certain prospects of researchability on communication by people. Permanent communication was traditionally made such by writings and audiovisual registrations and therefore rendered researchable only if archived. In time, this has produced a distinction, within society, between a *cared-for*-semantic (Luhmann 1996) (researchable)—or rather a semantic passed through mechanisms of production and distribution of cultural industry—and a semantic that, although created from within communication practices, could not have a sedimentation because, even when permanent, it was not visible and therefore not researchable. In this sense, much of our analysis on society was based, inevitably, on the analysis of a semantic that was necessarily processed by the cultural industries production system: books, films, newspaper articles, journalist reports, comics, magazines, photo stories.

Today instead the digital scenario offers the possibility of moving our research in *uncared-for*-semantic: that is diffused interpersonal communication, those UGC that have a sedimentation into the web, becoming both visible and researchable (Boccia Artieri 2012).

The point therefore is not only if and how to modify our research methods to collect a new kind of data but also how to define the nature of this social data.

From a methodological point of view, the user-generated content can be defined as qualitative data produced spontaneously by users, without any form of stimulation from the researcher, for an unknown audience.

From the point of view of data availability in SNSs, we are dealing with a quantity of data that was traditionally inaccessible to sociological analysis both for quality and quantity.

For a long time, the tradition of analysis of social relations within digital networks has always preferred a qualitative approach strongly tied to the ethnographical form of research (Boyd and Ellison 2007). This approach has certainly allowed a deeper level of understanding of the practices but has always found difficulty in describing macro-phenomena without the intermediation of single experiences. In recent years, traditional research approaches have been teamed with methods that are often defined as data driven: These approaches begin from a group of quantitatively relevant data to then identify paths of qualitative analysis (Magnani et al. 2010; Bruns et al. 2011).

2 Computational Social Sciences and Big Data: New Research Questions

This area of research defined as computational social science deals with the problem of Big Data and the shift it creates in how we think about our research.

As argued by Lazer et al., computational social science offers “the capacity to collect and analyze data with an unprecedented breadth and depth and scale” (2009, p. 722). But the issue of Big Data has not only to do with the detection and analysis of a large amount of data, but also to a *computational turn* in thought and research (Boyd and Crawford 2012). The tools we use to collect and analyze the data are indeed not neutral and require to rethink frame of our research and have an effect on the shaping of the theories that we formulate.

The wealth we obtain in this research prospective is counterbalanced by the problems that this approach involves in the difficult balance between the collection and quantitative data processing and interpretation of the observable results.

According to dana Boyd and Kate Crawford (2012):

> Due to efforts of mine and aggregate data, Big Data is fundamentally networked. Its value comes from the patterns that can be derived by making connections between pieces of data, about an individual, about individuals in relation to others, about groups of people, or simply about the structure of information itself.

This networked, connected, and relational dimension of Big Data is such that its possibility to visualize and recognize patterns derived from the connections (Srivastava et al. 2000) produces a reality of the research rather than just limited to representing answers to research questions.

We have also the risk of an apophenia effect: It is possible to observe emerging connections of every type. In this sense we must be supported by good research questions and good theories of reference in order to avoid getting lost in irrelevant streams where that which is technically visible is not socially relevant. For this, it is necessary to ask the correct questions and allow them to drive the research.

Take, for example, the attempt to predict the electoral outcome by measuring the quantitative data of conversations about political candidates on Social Network

Sites (Boccia Artieri 2013). As a paper of the Pew Research Center about Twitter and political campaign makes clear (Mitchell and Hitlin 2013):

> The reaction on Twitter to major political events and policy decisions often differs a great deal from public opinion as measured by surveys […]At times the Twitter conversation is more liberal than survey responses, while at other times it is more conservative. Often it is the overall negativity that stands out. Much of the difference may have to do with both the narrow sliver of the public represented on Twitter as well as who among that slice chose to take part in any one conversation.

Or take, for example, marketing, with its predictive studies on purchasing behavior and preferences of people gathered from the monitoring of what happens on social networks. These analyses allow the explanatory power to the design of algorithms rather than leading questions with purposeful (Goel et al. 2010).

We must underline that from a statistical point of view, research using Big Data has a problematic relationship with respect to representativeness (Cheueng 2012). If it is possible to trace, for example, the account name of a user is more difficult to have other elements (gender, age, education level, etc.). This fact makes it difficult to ensure representativeness. Building the representativeness requires the use of predictive models, thus to build complex algorithms. This solution at the moment remains mostly a horizon of possibilities of existing research (Daas et al. 2013).

3 Big Data: Critical Issues

To observe this reality, in particular inside social network sites, means to take into consideration a complex scenario, which is characterized by different critical issues.

3.1 The Contingency of the UGC: The Dimension of Time Is Relevant

A first issue that we have to deal with is the contingency of the dataset that we extract from social network sites. This requires us to face the dimension of time and the evolution of the network that we are observing. Much research tends to squash the analysis on a photograph of the data collected, analyzing it according to a cumulative logic that does not consider the longitudinal nature of the online conversations.

If for example we take the final dataset of tweets, replies and re-tweets of the users gathered around a specific #hashtag we obtain a photograph of the networks that eliminate the timing of events and squashes time into a sort of present continuous.

Let us try, for example, to observe all the data gathered around the hashtag "#earthquake" on Twitter as we did for the research on the propagation of content regarding the earthquake which hits the Emilia Romagna Region in Italy on 2012 with the first quake at 4.03 in the morning. This is a research which experiments a collection of data in real time on Twitter which is part of the Progetto di Rilevanza Nazionale 2009 (National Project of Importance 2009) that was financed by the Ministry of Universities and which I coordinate (Boccia Artieri et al. 2012).

By examining the data as a whole, it is possible to observe the degree of centrality of users associated with mass media outlets (either journalists or institutional accounts). If instead we observe the data in a gradual way, for example, the first 5 h, we find 95% of content was produced by simple users and that the presence of the mass media only arrived much later.

The first phase is obviously characterized by content with the highest percentage of propagation concerning the information about the event given by direct witnesses and the description of emotions and anxiety due to lack of information. Then, the informative level increases due to two phenomena: on the one hand, the beginning of production of information in other sites—not media but, for example, national seismological institutes—that are searched and shared in Twitter by users; and on the other, by the users themselves who, after the initial confusion, become the first reporters in the field. Having new instruments of communication and making the sharing of photographs and videos of the zones struck easier, therefore, many of these tweets are witness reports from the users giving detailed reports of the situation.

We can therefore observe the ReTweet chain of the second event when the rescue operations were already operating. The most prominent account is #INGVterremoti, the official Twitter account of the National Institute for Geology and Volcanology, that gives technical information of every seismic event happening in Italy. What is interesting to point out is the absence of public authorities of any kind (no local or central government or other state agencies) and the minor presence of mainstream news (@fattoquotidiano a well-known Italian newspaper seems to be an exception). The propagation of information about how to handle the crisis has been spread by minor users and local Twitter celebrities who used their followers as a propagation resource.

By observing the data in a gradual way, our interpretation about the propagation of the information in Twitter changes.

3.2 The Architecture of the Platforms Counts

A second issue that we have to deal with is that regarding the types of online conversations that we observe and we can summarize claiming that: platforms count and are not neutral.

From a research point of view, we must not forget that the characteristics of the networks studied with these methods are deeply influenced by the type of platform used to generate and observe them. Some platforms, for instance, are based on an asymmetric relationship among those users implying user A can follow B without B needing to do the same (this is the case with Twitter or Instagram, for example). On the contrary, other platforms (such as LinkedIn or, previously, Facebook) require a mutual acceptance of the friendship relation. It is just an example of how different platforms can originate different networks and, above all, how the platform, which in this case becomes also an observation tool, can influence the system we are displaying and can influence our conclusions from it.

From the point of view of the policies that are behind the platform, we find "tensions inherent in their service: between user-generated and commercially produced content, between cultivating community and serving up advertising, between intervening in the delivery of content and remaining neutral" (Gillespie 2010).

This characteristic also limits the boundaries of what we can or cannot observe on social network sites. To give a well-known example, we can consider certain activities on Twitter of purchasing of followers for brand accounts and celebrities and how this can unbalance the design of the network that we observe in terms of knots involved and the possibility of diffusion of the content.

Another crucial point when we talk about the influence of platforms on our research deals with the different privacy policies: these policies limit the possibility of collecting data and involves ethical concerns.

Not all content published on Internet is in fact completely public, and in social networks, the users select ever increasingly small circles of contacts with whom to share specific content—let us think of Google+ or Facebook—which make this content permanent but not freely searchable. For example for this reason, the sampling of FB is problematic due (1) to the lack of public timeline and (2) to privacy settings.

For this reason to analyze the Facebook profiles of Italians in our PRIN 2009 research, we developed an app to circulate and install using logic of gamification. Once the user accepted the terms and conditions of use of the app, the following data are stored in our database:

- Profile data, when indicated (religion, age, etc.)
- Posts—and their relative privacy settings—the questions and the notes, friend's comments,
- Friend's names (social network) and the lists in which the user has classified their friends.

This makes the approach extremely explorative because it is not possible to have control over some representatives but only to collect useful elements as a background to compare the qualitative analysis carried out by in-depth interviews. From the data collected, behavioral sets and topics can emerge: It is possible to build a map of relations and content to compare with the results of the interviews.

3.3 Making Sense of Network

The science of the networks must be contextualized considering the type of network that we are dealing with. One thing are the "topical network" produced by aggregated conversations around Twitter hashtags that take place outside of the Twitter network made of followers and friends; another are the "network structure" of followers and friends that shows a certain level of stability. These are not the same kind of networks.

The topical network, as we know, is built around a specific event and connects users via participation of the same theme highlighted by the chosen #hashtag. The strength of ties of this network will be necessarily different from that of a structured network, and the meaning of sharing of content or a reply to a user must consider the nature of the topical network itself but also of the structural network of the user. For example, we can compare the structural network of a user who participates on a topical network with the network as purely communicative (without the necessity of following or followers) which develops around a #hashtag.

Let us take the case of analysis of construction of connected networks beginning from participation, as audience, to the same media event. The question that we can ask is: In what way does the fruition of a television program according to ever increasing practice of social television modify social relationships that users have in social network sites?

The research "Conversation Practices and Network Structure in Twitter" (Rossi and Magnani 2012) monitored a specific hashtag (#XF5) that was used to tweet about the fifth edition of X-Factor Italy. The main results of the research show that being an active author on a widely discussed topic seems to increase the chances of getting new followers: Every user in the sample of the research obtained on average more than seven new followers after being active on the #XF5 hashtag. Obviously, it is not enough to be active in producing tweets: We have a positive connection with the acquisition of followers when tweets produce social relationships, that is, when they develop replies and retweets.

The finding show, as I pointed out earlier, how important is the variable "time" in structuring online social networks. By analyzing users who had acquired followers after a month, the researchers discovered that almost all the relationships built had been deleted. We can say that in time, this kind of relationship which was formed based on a momentary sharing did not generally manage to develop into mutual relationships of reading, reply, and retweet. These relationships are the first to be subjected to unfollow practices.

3.4 Deep Data and the Context of the Facts

A last issue that we have to deal with is about how, in an era of the Big Data, we can contemplate research data that relies more on the context of creation than on the volume and variety of sources. And how to go deeper into “sense making” of the big volume of data collected.

After all, Big Data is the collection of those thin data that we monitor in our daily lives by tracing actions and behaviors. The location and time of our morning run or the time spent sleeping, the relationships we have with others, the kind of music we like to listen to or the books we love to read, all those are data that are monitored by cookies on the Web sites that we frequent, from the apps related to wellness, fitness, diet; or from the Fitbit we are wearing or from our smartphones with their GPS on, etc.

They are essential data about us, but they only partially tell things about us. They do not represent our experience of the world and the sense that we put into the things we do. To really understand the aspects of our experience, we need Deep Data. Deep Data capture not just facts, but the context of facts. Rather than try to understand us simply basing the comprehension on “what we do”—as in the case of Big Data—Deep Data seeks to understand us in terms of “how we relate to” the many different worlds we inhabit. They give back the sense of the complexity of the actions and behaviors of our everyday lives.

Addressing the problem of depth and context analysis means using Big Data to identify patterns and signals to be qualitatively checked with our methodological approaches. And perhaps this is the most significant contribution in data science that social sciences can and should give.

4 Conclusions

We are faced with an experimental methodology that we still have to fully explore and whose criticisms must be carefully considered. The main platforms of reference tend to change rapidly in time redesigning the environments and forming different communication activities. Also, the policies regarding the treatment of data evolve with the evolution of the functionality offered by the service itself. As computational social scientists, we have to follow a road in which, as I have tried to explain, we have to learn to ask the right questions with a transdisciplinary view. Questions which have to consider the nature of the platforms that host the communication environments that we observe. Questions which consider the specific nature of the network we have chosen to observe and that are able to read their evolution over time. Questions which consider the relation between Big Data and Deep Data, to capture the social context.

It is a difficult task that we are commencing, discussing the many doubts, and sharing the results. It appears, however, certain that social media open new roads to the scholars of society that show great potential of development and maybe in the direction of a true and real paradigm shift in social sciences. At the same time, it also appears clear that the necessary ability to take full advantage of the potential offered by this new field requires the collaboration of scholars originating from various fields of knowledge. Computational social sciences, if they ever exist, will be the meeting place of scholars from different disciplinary backgrounds and not a new independent field of knowledge.

References

Boccia Artieri, G. (2012). Productive publics and transmedia participation. *Participations. Journal of Audience & Reception Studies*, *9*(2), 448–468.

Boccia Artieri, G. (2013). Un tweet non fa l'elettore. In I. Diamanti (Ed.), *Un salto nel voto. Ritratto politico dell'Italia di oggi* (pp. 167–182). Roma-Bari: Laterza.

Boccia Artieri, G., Giglietto, F., & Rossi, L. (2012). #terremoto! L'uso di Twitter durante il terremoto tra testimonianza, propagazione e commento. https://snsitalia.wordpress.com/2012/05/24/terremoto-luso-di-twitter-durante-il-terremoto-tra-testimonianza-propagazione-e-commento/.

Boyd, D. (2008). Taken out of context: American teen sociality in networked public. Ph.D. dissertation, School of Information, University of California, Berkeley.

Boyd, d., & Ellison, N. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. doi:10.1111/j.1083-6101.2007.00393.x.

Boyd, d., & Crawford, K. (2012). *Six provocations for big data*. Paper presented at Oxford Internet Institute's "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. Retrieved April 30, 2014, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431.

Bruns, A., Burgess, J., Highfiled, T., Kirchhoff, L., & Nicolai, T. (2011). Mapping the Australian networked public sphere. *Social Science Computer Review, 29*(3), 277–287. doi:10.1177/0894439310382507.

Cheung, P. (2012). *Big data, official statistics and social science research: Emerging data challenges*. Presentation at the December 19th World Bank meeting, Washington. http://www.worldbank.org/wb/Big-data-pc-2012-12-12.pdf.

Daas, P. J. H., Puts, M. J., Buelens, B., et al. (2013). *Big data and official statistics*. Paper for the 2013 NTTS Conference, Brussels, Belgium. Retrieved March 5–7, from http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_76.pdf.

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society*, *12*(3), 347–364.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences, 7*(41), 17486–17490.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., et al. (2009). Social science. Computational social science. *Science, 323*(5915), 721–7233. doi:10.1126/science.1167742.

Luhmann, N. (1996). *Die Realität der Massenmedien*. Opladen: Westdeutscher Verlag.

Magnani, M., Montesi, D., & Rossi, L. (2010). Friendfeed breaking news: Death of a public figure. In: *Second IEEE International Conference on Social Computing*. Los Alamitos, USA: IEEE computer Society, 528–533.

Mitchell, A., & Hitlin, P. (2013). Twitter reaction to events often at odds with overall public opinion. http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/.

Rossi, L., & Magnani, M. (2012). Conversation practices and network structure in Twitter. In: *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media* (pp. 563–566).

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining, 1*(2), 12–23.

Governing by Data: Some Considerations on the Role of Learning Analytics in Education

Rosanna De Rosa

Abstract E-governance is growing in interest as a technique of decision making in the public domain. The technological revolution is, in fact, enabling governments to use a great variety of digital tools and data to manage more effectively all phases of the policy cycle process. Although still in its infancy, the use of big data is becoming a core element of the emergence of e-governance as a form of evidence-based policy making. We are assisting to the emergence of a form of *governing by data*, where governments will take more and more sensitive decisions through data mining systems. Education is one of the policy fields where the use of big data such as learning analytics is becoming crucial for funds allocation, institutional accountability, and programs review. This article briefly explores the scientific debate about analytics highlighting the role they should play in the educational datascape.

1 Introduction

The concept of big data is attributed to Laney (2001), who described their main characteristics as its size (volume), speed (velocity), and its varying shape (variety). More recently, the definition has been revisited with the addition of one more characteristic: Quality (Veracity), meaning that the inclusion of external and heterogeneous data—even though big—still raises questions about the accuracy and completeness of datasets (Chen et al. 2014). More in general, different levels of analytics combine to form a set of data that can provide useful pointers for making organizational decision making more effective. Yet, their use for policy setting is strictly related to the nature of data, statements, algorithms, and—of course—data interpretation.

R. De Rosa (✉)
Department of Social Sciences, University of Naples Federico II,
Vico Monte Di Pietà 1, Naples, Italy
e-mail: rderosa@unina.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_7

In the educational field, big data tends to coincide with the analytics framework in all its variation—learning, institutional, academic—and specification—retrospective, real time, and predictive. A clear conceptual framework on the definition of analytics in their different operational context is offered by van Barneveld et al. (2012) that distinguishes learning effectiveness from operational excellence: «with the latter referring to the metrics that provide evidence of how the training/learning organization is aligning with and meeting the goals of the broader organization. Learning analytics in the academic domain is instead focused on the learner, gathering data from course management and student information systems in order to manage student success, including early warning processes where a need for interventions may be warranted» (p. 6).

The introduction of various educational software systems dramatically changed the process of educational delivery for both distance and on-campus modes of instruction. Teaching in the digital era is, in fact, characterized by a predominance of complexity. It is no longer being perceived as part of the supply demand production chain. At the same time, learning becomes an interconnected experience where the learner alternates between the physical and the digital space, between production and consumption, and between formal and informal learning. Also the new concept of MOOC and MOOC delivery platforms offers the potential to reshape important aspects of the policy-making environment in the educational field, with analytics playing a new role for policy choices.

This explains why the set of data (learning analytics) provided by online learning tracking systems (including logs, quizzes, and social network analysis) are growing in popularity as a way of giving face-validity to learning outcomes in the digital world. It is defined as

> an educational application of web analytics aimed at learner profiling, a process of gathering and analyzing details of individual student interactions in online learning activities (Horizon Report 2016) (p. 38),

learning analytics aim at building better pedagogies, empowering active learning, targeting at-risk student populations, and assessing factors affecting completion and student success.

Due to the profound implications of analytics, it becomes urgent to define a clear policy framework, starting from the definition of a shared research agenda. In a situation where limited financial and human resources is a common ground on which higher education institutions are trying to reform themselves, it is not a surprise if analytics are becoming a core issue for policy makers and stakeholders, acting for scaling up data for a more *efficient* education system. But what does *efficient* learning really means? And how can data-driven decisions really help to cope with this increase in complexity and fast changing social environment?

In the next paragraphs, we will present some arguments in support of data use.

2 The Transformative Power of Data

Approaching big data in education from the business perspective is the best way to understand their disruptive potential. It is also helpful to understand the reason why strong resistances are usually raised from diverse points of the system. Analytics has been, in fact, recognized as one of the most relevant and growing submarket fields in education. Visiongain (2015) considers analytics as the foremost reason universities are offering open MOOCs as big data technologies allow to collect and analyze valuable data derived from online education, to identify strengths and weaknesses and to reach conclusions that will raise the overall value of the institution. A submarket estimated to grow from 27.5 billion dollars in 2015 to 109.3 in 2020 (CAGR of 36.1%). According to this approach, the WICHE Cooperative for Educational Technologies (WCET), using data accumulated by the U.S. Department of Education, can definitively state that distance education is no longer an institutional accessory, highlighting the current state of the art in the distance education industry, which has passed from 1.6 million students in 2002 to 5.8 in 2014.

In McKinsey Institute Quarterly titled *Are you ready for the era of 'Big Data'* Brown et al. (2011), it is clearly stated that big data will become a new type of corporate asset representing a key basis for competition. The McKinsey report also suggests how big data potential can be explored and exploited to create potentially disruptive business models by transforming processes, altering corporate ecosystems, and facilitating innovation. Although the McKinsey narrative is not expressly devoted to education, the framing of education as an economic activity is behind the logic of using big data for business intelligence in education, especially in such areas as outreach and advertising, enrollment, management, personnel recruitment, and fundraising (Clow 2013). Moreover, as already happened in several human activities based on the *productive* paradigm, investments in education are increasingly connected to data-driven decision making, thus shifting the focus from student success and innovative pedagogies to institutional effectiveness and return of investments (ROI) strategies. In some cases—such as the online education—this implies an almost direct correlation between the success of the learning experience, and the political and social approval of financial investing in digital learning.

Following this economic framework, academic institutions are becoming aware of the potentiality of analytics starting to test technologies, data mining processes, visualization modeling, and dashboards. They are following a *natural* path to move analytics from hindsight to foresight, from description to prediction through a diagnosis of what to change in teaching methods and organizations. The next paragraphs are organized around three main spheres—the systems of measurement, influence, and evidence—that all together shape the special datascape for policy making in education (Fig. 1).

It tries to highlight the role of pedagogies, the space reserved to software development, and the type of actors involved in decision-making processes.

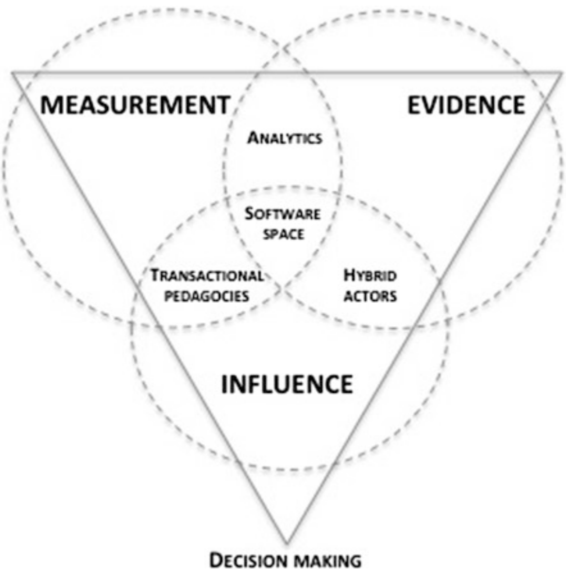


Fig. 1 Datascape in the education policy making

3 System of Measurement

Leveraging Analytics in Community Colleges is an interesting Educause guide (Educause Review 2015) providing a literature review, list of definitions, and resources to community college leaders. It distinguishes different types of analytics related to different use/purpose of data: *academic analytics*—data collected to support operational and financial decisions—*learning analytics*—feedback about teaching and learning performances—and *predictive analytics,* aiming at identifying trends to forecast the future on academic, operational, and financial plans (Long and Siemens 2011; Ferguson 2012a). Feedback loops, policy indications, and operational knowledge are expected from data use; meantime, several case studies to be inspired by are already available at any level [see LAK Conference 2016]. The impact of analytics on social organization is very much related to the chosen option.

3.1 Description

A few European projects are piloting learning analytics in the online education context. This is the case of the Emma Project (http://www.europeanproject.eu) which is experimenting with data at user and platform level to raise the awareness of the MOOCs' participants' learning activities as well as to provide feedback for MOOC instructors about their course design. This with a dual perspective: real time and retrospective analytics. This kind of approach, although interesting in helping instructional designers to create online courses with higher retention and completion rate, focuses mainly on description of learning behavior with little capacity to distinguish prospective students and/or offer adaptive content. However, in the learner's perspective, the possibility to compare themselves with the performances of the online classroom is valuable even though it can exercise a form of pressure toward conformity. Investigating the impact of data on the teacher and student means, however, to identify not only strengths and weaknesses of analytics tools and practices, but also their role for social control and/or policy effectiveness.

3.2 Prediction/Prescription

The Predictive Analytics Reporting (PAR) appears one of the most large scale projects so far. PAR is a framework (PARframework.org) adopted by 35 US academic institutions and created with the participation of 350 campuses with millions of students' performance already processed. It uses descriptive, inferential, and predictive analyses to create benchmarks, institutional predictive models, to map, and measure student interventions that should have direct positive impact on behaviors correlated with success. Using data mining on a federate datasets of millions of de-identified student records, this framework should be able to identify those variables affecting student achievement and performance. Acting as a non-profit, multi-institutional collaborative venture, the PAR framework focused on leveraging common data definitions and predictive analytics in the service of student success using common data definitions for core measures across institutions to seek patterns of student loss and success. However, Ellen Wagner, Chief Strategy Officer for Predictive Analytics Reporting (PAR) Framework, in her presentation at Online Educa Berlin 2015, introduced the concept of data as a *meme*, focusing on the dangers of «naïve or nefarious uses of data to restrict access or to punish», opening up the door to more reflection on the *hidden curriculum effect.*

3.3 Accountability

The Research Assessment Exercises in Italy—known as Research Quality Evaluation (VQR)—is an assessment system where data are used mainly for system accountability by stakeholders in order to reorganize the academic world, reducing costs, increasing productivity, and allocating human resources and funds in a more efficient way. Representing the VQR as only a face of the academic work—which concerns a complex set of interconnected activities, from teaching to community networking, from projecting to expert support in decision-making processes, as well as institutional organization, dissemination, and social work—it is not a surprise that academic staff often refused to legitimize the VQR exercise as representing only a 'single dimension' of academic work. System accountability is, in fact, strongly rooted in the *datascape*, with a substantial impact at an organizational and academic level. Yet, the system—because of its inherent push toward a flat standardization of metrics and productivity assessment—is producing ambiguous and—sometimes—non acceptable results, creating struggles and discontent in the academic staff. Although the Italian Minister of Education has recently created a working group on big data in order to support system decision making, a more complex data-driven profile of academic work is still far from being realized, and—overall—socially recognized. However, it seems that the more we claim for a multimodal approach, the more is the space of opportunity for the application of Educational Data Mining (EDM), focused on prediction, clustering, data mining, and distillation of data for human judgment [see LACE Project 2014]. An holistic approach to the analytics field seems then the only solution to offer a social vision of its use, putting the development of the human being (learners, teachers, researchers etc.) at the very core of policy making.

4 Culture of Evidence

The McKinsey Report (Manyika et al. 2013) considers as the main challenge of this new century the possibility to leverage the transformative power of big data to create transparency, to enable experimentation and to discover new needs. While, from the institutional policy perspective, it helps to expose variability and improve performance segmenting populations, to customize actions and replace/support human decision making with automated algorithms, and, finally, to support innovation in new business models, products, and services. In other words, big data and analytics require a change of paradigm and the foundation of a new culture of data, available at any level of the societal interest. Only an intelligent management accompanied by organizational capacities and institutional commitment can offer such a possibility helping institutions to prevent bias associated with the emergence of a new field.

4.1 Short Circuits

The national pupil database, established in 2002 by the UK government, is a central policy instrument of educational governance used to translate massive data into actionable policy indications as well as school performance tables accessible to parents and media. In Italy, taken into account the PISA test, the education policy program—known as the National Plan for the Digital School (PNSD)—has activated a plethora of actors, agencies, platforms, and networks acting as a learning ecosystem in which schools have no longer the monopoly of teaching but receive support from different education providers. Activated by evidence-based reasoning (i.e., digital skills in education and performance in STEM domain), this learning ecosystem aims at innovating pedagogies, engaging teachers in reinventing processes, and introducing new assessment forms.

Data governance is usually—or at least should be—oriented to understand why things happen, what are the current trends, their impacts, and how to orchestrate the appropriate answers to meet the organization's goals. In other words, it is positioned

> «to short-circuit existing educational data practices, enabling data and feedback to flow synchronously and recursively within the pedagogic apparatus of the classroom itself» (Norris and Baer 2013).

It requires an efficient and complex combination of data stewardship, reporting, query, and analytics tools. Only this combination can help policy makers to monitor events and take strategic decisions on both short and long terms.

4.2 Collateral Risks

Indication about how students are performing in a comparative way could, more or less implicitly, act to suggest that some students are not good enough to keep studying—or to deserve our financial investment —and, so, it may reproduce social inequalities. Scholars called it the *hidden curriculum effect* (Edwards 2015).

As researchers, we do not take too seriously the social implications of software in education, and how they are able to shape our lives. Yet the discussion about the verifiability of the *hidden curriculum effect* needs to be at very heart of the academic discussion and political debate in order to evaluate both evidences (issue representation by data) and results (output and outcomes).

As the interest in evidence-based policy making is increasing and online education is gaining momentum, it is imperative to pose the issue of technology embedded into education as an urgent research question, since we cannot consider computer technologies simply as a tool by which learning is delivered (Edwards 2015). Hence, a genuine *culture of evidence* should be never untangled by a deep *culture of social inquiry*.

4.3 Methods

Experts argue that the most valuable plus of analytics is that they can help managers distinguish *causation* from mere *correlation*, thus reducing the variability of outcomes while improving financial and institutional performances. The risk for socio- and techno-determinism is, instead, what social sciences really want to avoid in order to provide complex multimodal explanations. It is for this reason that the learning analytics community need to build stronger connections with the learning sciences, to develop methods of working with a wider range of datasets as well as to avoid simplistic data interpretation and misuse. Ferguson (2012) indicates in this link a way to ensure the optimization of learning environments under the guidance of a clear set of ethical guidelines.

However, to really exploit the data universe and optimize methods and techniques, social sciences needs a number of devoted researchers, people with deep analytical skills and data-driven mind-set, but also a number of scientists and practitioners able to use and transform the information based on data into actionable knowledge. This implies creating a genuine culture of evidences inserting it into institutional and organizational practice. It implies also the need to explore and assess the qualitative value of quantitative data.

5 System of Influence

Even though the *what* and the *how* of analytics have been explored enough by the specialized literature, the *who* still remains a question mark. Who is in charge for data translation into policy, with what kind of expertise, and goals in mind, is probably a crucial research question that it is not possible to avoid any more.

5.1 Hybrid Actors

The British Education Endowment Foundation (EFF) is part of The What Works center acting, with the govern legitimation, to share research with local decision makers. EEF provides independent and accessible information teaching and learning toolkits summarizing educational research from the UK and around the world. EFF participates also to the Alliance for Useful Evidence—an open access and wide network of individuals and organizations interested in promoting useful evidences in decision making across social policies. This case works quite well as an example of what Ben Williamson, in *New governing experts in education: Policy labs, self learning software, and transactional pedagogies* (Williamson 2014), sees as the emergence of a hybrid actor—in between think tank, R&D, and social enterprises—that produces material able to transform education by making it «problematic, thinkable, intelligible, and hence practicable in new ways» (p.2). Such material is in large part derived from data-driven technologies to support forms of self-regulated public policy where the learners are considered a *calculable* governing resource. What is evident is the shift from the formal organs of government toward a larger network of actors playing their role on different basis (i.e., commercial and nongovernmental) (Lynch 2015).

5.2 *Software Space*

The Pearson Learning Curve is an example of a project that offers country indexes, country profiles, and comparable data set and visualization in education by time and outputs. A project explicitly devoted «to help influence education policy and practices, at local, regional and national levels» (Williamson 2016a).

Clearly, from different points of the social systems, there is a claim to expand the capacity of organizations to make sense of complexity. This goal, however, is strongly related to the creation of an expert software system for governing education. About that Lynch (2015) in his book, *The Hidden Role of Software in Education*, explored how a new kind of 'software space' (code, algorithms, dataset) is joining the 'political space' of educational governance, influencing the 'practice space' of the classroom. In that sense, the 'software' space is playing an agency role among a wide network of actors.

5.3 *Transactional Pedagogies*

Williamson (2016b) suggests that the role of data and software is much more than offering guidance to stakeholders throughout the decision-making process. What he finds particularly worrying is, in fact, the translation process of problems, ideas, practices into "inscription" devices such as visualization, infographic, images that all contribute to generate a representation of a «governable education system» (p.11). The conclusions highlight the emergence of transactional pedagogies and transactional policies, both data-driven, as ideal progress of the 'knowing capitalism.'

6 Are Big Data Empirical Evidences Only? Some Final Concerns

The aforementioned Educause report raised warnings about student profiling techniques, affordability, and misuse of data, privacy issues as well as business-like practices, data ownership, and their exploitation. This criticism is only a first step toward a full social awareness about the ontological implications of big data concerning, for example, the pressure toward conformity, dataveillance, and the increased power of microstructure and macrostructure through data governance. The scientific debate, however, has a plurality of voices.

In his seminal book *The Philosophy of Software, Code and Mediation in the Digital Age* (Berry 2011), Berry defines as *datascape* the computational narrative of the subject represented by all data concerning his activity streams; a datascape that can influence very deeply our present and future lives. Boyd and Crawford (2012) claim that big data are represented as a higher form of intelligence able to generate insights with «the aura of truth, objectivity, and accuracy» (p. 663) but they are not, since numbers do not speak for themselves. On the same direction, Kirschner contrasted the use «to look at data we have and not at data we need» to make inferences [LAK2016 conference].

Concerns were raised also during a debate at Open Education Berlin (OEB2015). Mayer-Schönberger, Professor of Internet Governance and Regulation at Oxford University's Internet Institute, linked the use of data to human progress to demonstrate how data has been and will be even more necessary. So Darrell West, author of *Digital Schools: How Technology Can Transform Education* (West 2013), considers the introduction of analytics at any level not only useful but also necessary since «schools face a situation where they need to improve the overall accountability of their operations» (p.9). For G. Siemens, one of the inventors of MOOCs, analytics is a cognitive process that makes data more manageable, enabling us to make sense of the world.

Even though one could try to balance pros and cons, the use of big data and analytics in education implies a dimension of social control never experienced with such a level of efficiency and pervasiveness. Adopting the science and technology studies perspective (STS), the governance system based on digital technologies must be considered as a policy instrumentations for social control, partial, or fictitious representations of that reality that one want to change. With the relevant implication of empowering software, data companies and agencies have to become dominant and stable partners for governing education.

References

Berry, D. (2011). *The Philosophy of software, code and mediation in the digital age* (p. 171) Palgrave MacMillan.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication and Society, 15*(5), 662–679.

Brown, B., Chui, M., & Manyika, J. (2011). *Are you ready for the era of 'Big Data'*. McKinsey Institute Quarterly October 2011.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications, 19*(2), 171–209.

Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education, 18*(6), 683–695.

Educause Review. (2015). Leveraging analytics in community colleges.

Edwards, R. (2015). Software and the hidden curriculum in digital education. *Pedagogy, Culture and Society, 23*(2), 265–279.

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*(5), 304–317.

Horizon Report. (2016). Higher education edition. The New Media Consortium:56.

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

Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *46*(5) 34.

Lynch, T. L. (2015). *The hidden role of software in educational research*. Routledge: 230.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. et al. (2013). *Big data: The next frontier for innovation, competition, and productivity* (p. 156) The McKinsey Global Institute Report.

Norris D. M. & Baer L. L. (2013). Building organizational capacity. *EDUCAUSE.*

van Barneveld, A., Arnold K. E., & Campbell J. P. (2012). Analytics in higher education: Establishing a common language. *EDUCAUSE, 2012*. (p. 2). ELI Paper.

Visiongain. (2015). *Massive open online course (MOOC) market 2015–2020* (p. 136) Report.

West, D. (2013). *Digital schools: How technology can transform education* (p. 159) Brookings Institution Press, 2012.

Williamson, B. (2014). New governing experts in education: Self-learning software, policy labs and transactional pedagogies. In T. Fenwick., E. Mangez & J. Ozga (Eds). *World yearbook of education 2014: Governing knowledge: Comparison, knowledge-based technologies and expertise in the regulation of education.* London: Routledge, pp. 218–231.

Williamson, B. (2016a). Digital methodologies of education governance: Pearson plc and the remediation of methods. *European Educational Research Journal, 15*(1), 34–53.

Williamson, B. (2016b). Digital education governance: An introduction. *European Educational Research Journal, 15*(1), 3–1.

Part II
Methods, Software and Data Architectures

Multiple Correspondence K-Means: Simultaneous Versus Sequential Approach for Dimension Reduction and Clustering

Mario Fordellone and Maurizio Vichi

Abstract In this work, a discrete model for clustering and a continuous factorial one for dimension reduction are simultaneously fitted to categorical data, with the aim of identifying the best partition of the objects, described by the best orthogonal linear combinations of the factors, according to the least-squares criterion. This new methodology named multiple correspondence *k*-means is a useful alternative to the Tandem Analysis in the case of categorical data. Then, this approach has a double objective: data reduction and synthesis, simultaneously in the direction of rows and columns of the data matrix.

1 Introduction

In the era of "*big data*", complex phenomena—representing reality in economic, social and many other fields—are frequently described by a large number of statistical units and variables. Researchers who have to deal with this abundance of information are often interested to explore and extract the relevant relationships by detecting a reduced set of prototype units and a reduced set of prototype latent variables, both representing the "*golden knowledge*" mined from the observed data. This dimensionality reduction of units and variables is frequently achieved through the application of two types of methodologies: a discrete classification method, producing hierarchical or non-hierarchical clustering and a latent model, creating factors. The two methodologies, generally are not independently applied. In fact, first, the factorial method is used to determine a reduced set of latent variables and then the clustering algorithm is computed on the achieved factors. This sequential strategy of analysis has been called Tandem Analysis (TA) by Arabie (1994). By applying first the factorial method, it is believed that all the relevant information regarding

M. Fordellone (✉) · M. Vichi
Sapienza University of Rome, Rome, Italy
e-mail: mario.fordellone@uniroma1.it

M. Vichi
e-mail: maurizio.vichi@uniroma1.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_8

the relationships of variables is selected by the factorial method, while the residual information represents noise that can be discarded. Then, the clustering of units completes the dimensionality reduction of data by producing prototype units generally described by centroids, that is, mean profiles of units belonging to clusters.

However, some authors have noted that TA in some situations cannot be reliable because the factorial models applied first may identify factors that do not necessarily include all the information on the clustering structure of units (Desarbo et al. 1991). In other terms, the factorial method may filter out some of the relevant information for the subsequent clustering. A solution to this problem is given by a methodology that includes the simultaneous detection of factors and clusters on the observed data. Many alternative methods combining cluster analysis and the search for a reduced set of factors have been proposed, focusing on factorial methods, multidimensional scaling or unfolding analysis and clustering (e.g. Heiser 1993; De Soete and Heiser 1993). De Soete and Carroll (1994) proposed an alternative to the TA procedure, named reduced *k*-means (RKM), which appeared to equal the earlier proposed projection pursuit clustering (PPC) (Bolton and Krzanowski 2012). RKM simultaneously searches for a clustering of objects, based on the *K*-means criterion (MacQueen 1967), and a dimension reduction of the variables, based on component analysis. However, this approach may fail to recover the clustering of objects when the data contain much variance in directions orthogonal to the subspace of the data in which the clusters reside (Timmerman et al. 2010). To solve this problem, Vichi and Kiers (2001) proposed the factorial *k*-means (FKM) model. FKM combines *K*-means cluster analysis with PCA, then finding the best subspace that best represents the clustering structure in the data. In other terms, FKM selects the most relevant variables by producing factors that best identify the clustering structure in the data. Both RKM and FKM proposals are good alternative to the TA in the case numeric variables have been considered.

When categorical (nominal) variables are observed, TA corresponds to apply first multiple correspondence analysis (MCA) and subsequently the *K*-means clustering on the achieved factors. As far as we know there are no studies that verify if this TA has the same problems observed for quantitative variables. Thus, the first aim of this paper is to discuss whether there are limits of the TA in the case of categorical data. The second and most relevant aim of the paper is to present a methodology, named multiple correspondence *k*-means (MCKM), for simultaneous dimension reduction and clustering in the case of categorical data. The work is structured as follows: in Sect. 2 a background on the sequential and simultaneous approaches is provided, showing an example where TA for categorical data fails to identify the correct clusters. This is a good motivating example that justifies the use of a simultaneous methodology. In Sect. 3, details on the MCKM model are shown; in Sect. 4, the alternative least-squares (ALS) algorithm is proposed for MCKM. In Sect. 5, the main theoretical and applied proprieties of the MCKM are discussed; and finally, in Sect. 6, application on a real benchmark data is given to show the characteristics of MCKM.

2 Statistics Background and Motivating Example

Let $\mathbf{X} = [x_{ij}]$ be a $N \times J$ data matrix corresponding to N units (objects) on which J categorical (nominal) variables have been observed. Tandem Analysis (TA) (Arabie 1994; Desarbo et al. 1991) is the statistical multivariate procedure that uses two methodologies: (i) a dimension reduction (factorial) method for finding a set of P factors (generally, $P < J$) better reconstructing the J observed variables (e.g. by using principal component analysis (PCA) or factor analysis (FA)); and (ii) a clustering method that partitions the N multivariate objects into K homogeneous and isolated clusters (for example by considering K-Means, or Gaussian mixture models). In TA, the factorial method is applied first to compute a matrix of component scores; then, the clustering method is applied, sequentially, on the component score matrix. The first methodology detects the maximal part of the total variance by using a reduced set of P components; while the second method maximizes the between variance of the total variance explained in the first analysis. Thus, the variance explained by the factorial method could not be all the between variance of the original variables necessary for the successive clustering methodology. Actually, it may happen that some noise masking the successive clustering could have been included in the P components. Vichi and Kiers (2001) show an instructive example where a data set formed by variables with a clustering structure, together with other variables without clustering structure (noise), but having high variance, has been considered. When TA is applied on this typology of data, the PCA generally explains also part of the nose data. These last tend to mask the observed clustering structure, and as a consequence, several units are misclassified.

If the J variables considered in the matrix $\mathbf{X}$ are categorical, then TA corresponds, usually, to the application of multiple correspondence analysis (MCA) and K-Means (KM), this last sequentially applied on the factors identified by MCA. The researcher may ask if this TA for the categorical variables has the same limits discussed for the quantitative case. Before considering this, let us first formalize TA in the categorical data case.

The MCA model can be written as

$$J^{1/2}\mathbf{JBL}^{1/2} = \mathbf{YA}' + \mathbf{E}_{MCA} , \tag{1}$$

where $\mathbf{Y} = J^{1/2}\mathbf{JBL}^{1/2}\mathbf{A}$ is the $N \times P$ score matrix of the MCA; $\mathbf{A}$ is the $J \times P$ column-wise orthonormal loadings matrix (i.e. $\mathbf{A}'\mathbf{A} = \mathbf{I}_P$); $J^{1/2}\mathbf{JBL}^{1/2} = \mathbf{X}$ is the centred data matrix corresponding to the J qualitative variables, with the binary block matrix $\mathbf{B} = [\mathbf{B}_1, \ldots, \mathbf{B}_j]$ formed by J indicator binary matrices $\mathbf{B}_j$ with elements $b_{ijm} = 1$ if the ith has assumed category m for variable j, $b_{ijm} = 0$ otherwise; $\mathbf{L} = diag(\mathbf{B}'\mathbf{1}_N)$; $\mathbf{J} = \mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N'$ is the idempotent centring matrix with $\mathbf{1}_N$ the N-dimensional vector of unitary elements.

The KM applied on the MCA score matrix $\hat{\mathbf{Y}} = J^{1/2}\mathbf{JBL}^{1/2}\hat{\mathbf{A}}$ can be written as

$$\hat{\mathbf{Y}} = \mathbf{U}\bar{\mathbf{Y}} + \mathbf{E}_{KM} , \tag{2}$$

where $\mathbf{U}$ is the $N \times K$ binary and row stochastic memberships matrix, i.e., $u_{ik} \in \{0, 1\}$ with $i = 1, \ldots, N$ and $k = 1, \ldots, K$ and $\mathbf{U}\mathbf{1}_K = \mathbf{1}_N$, identifying a partition of objects and $\bar{\mathbf{Y}}$ is the $K \times P$ corresponding centroid matrix in the P-dimensional space. Note that $\mathbf{Y} = \mathbf{X}\mathbf{A}$, while $\hat{\mathbf{Y}} = \bar{\mathbf{X}}\mathbf{A}$. Finally, $\mathbf{E}_M CA$ and $\mathbf{E}_K M$ are the $N \times J$ error matrices of MCA and KM, respectively.

The least-squares (LS) estimation of model (1) corresponds to minimize the loss function

$$\begin{cases} ||J^{1/2}\mathbf{J}\mathbf{B}\mathbf{L}^{1/2} - \mathbf{Y}\mathbf{A}'||^2 \underset{\mathbf{A}}{\rightarrow} min \\ \mathbf{A}'\mathbf{A} = \mathbf{I}_P \\ \mathbf{Y} = J^{1/2}\mathbf{J}\mathbf{B}\mathbf{L}^{1/2} \end{cases}, \tag{3}$$

while LS estimation of model (2) relates to minimize the loss function

$$\begin{cases} ||\hat{\mathbf{Y}} - \mathbf{U}\bar{\mathbf{Y}}||^2 \underset{\mathbf{U},\bar{\mathbf{Y}}}{\rightarrow} min \\ \mathbf{U} \in \{0, 1\} \\ \mathbf{U}\mathbf{1}_K = \mathbf{1}_N \end{cases}, \tag{4}$$

Thus, given the LS estimates $\hat{\mathbf{A}}$, $\hat{\mathbf{U}}$, $\hat{\hat{\mathbf{Y}}}$ of MCA and KM and considering $\mathbf{Y} = J^{1/2}\mathbf{J}\mathbf{B}\mathbf{L}^{1/2}\hat{\mathbf{A}}$, the TA procedure has an overall objective function equal to the sum (or mean) of the two objective functions of MCA and KM; formally,

$$f(\hat{\mathbf{Y}}, \hat{\mathbf{A}}, \hat{\mathbf{U}}, \hat{\hat{\mathbf{Y}}}) = \frac{1}{2}\left(||J^{1/2}\mathbf{J}\mathbf{B}\mathbf{L}^{1/2} - \hat{\mathbf{Y}}\hat{\mathbf{A}}'||^2 + ||\hat{\mathbf{Y}} - \hat{\mathbf{U}}\hat{\hat{\mathbf{Y}}}||^2 \right). \tag{5}$$

Therefore, TA is the procedure that optimizes sequentially the two objective functions of MCA and KM, which loss can be summarized by (5). However, we now show with an example that this sequential estimation has some limits similar to those evidenced in the quantitative case. In Fig. 1, the heat map of the data matrix of 90 units according to 6 qualitative categorical variables, each one with 9 categories is shown.

This is a synthetic data set formed by considering multinomial distributions. The first two variables are a mixture of three multinomial distributions with values from 1 to 3, from 4 to 6 and from 7 to 9, respectively, thus defining three clusters of units, each one with equal size (30 units). The other four variables are multinomial distributions with values from 1 to 9 with equal probabilities; thus, these do not define clusters of units. We suppose that this is an example of a simulated data set of 90 customers who have expressed their preferences on 6 products on the basis of a Likert scale from 1 (like extremely) to 9 (dislike extremely), passing through 5 (neither like nor dislike). The heat map in Fig. 1 is a graphical representation of data where the individual values contained in the matrix are represented as different levels of grey from white (value 1) to black (value 9) (1 like extremely, 2 like very much, 3 like moderately, 4 like slightly, 5 neither like nor dislike, 6 dislike slightly, 7 dislike moderately, 8 dislike very much, 9 dislike extremely). By examining the columns of the heat map (corresponding to products), it can be confirmed that the first two (products

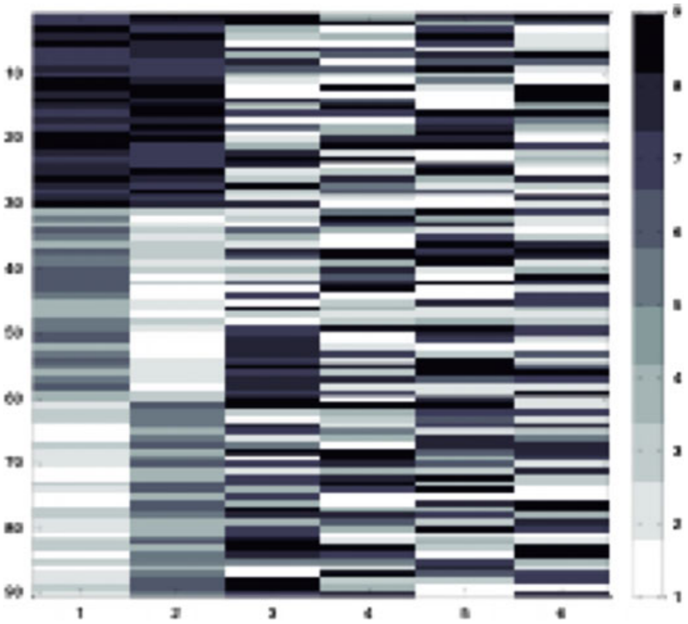


Fig. 1 Heat map of the 90 × 6 categorical variables with 9 categories for each variable

A, B) have a well defined clustering structure. In fact, the first 30 customers dislike (moderately, very much and extremely), the two products having chosen attributes from 7 to 9, for both products. Customers from 31 to 60 having values from 4 to 6 and from 1 to 3, for the first and second column, respectively, are almost neutral on the first product (like slightly, nether like nor dislike, dislike slightly), but they like the second product (extremely, very much or moderately). Finally, customers from 61 to 90 have values from 1 to 3 and from 4 to 6 in the first and second column, respectively; thus, they like the first product and are substantially neutral for the second. For the other four products (C, D, E, F), the 90 customers do not show a systematic clustering pattern with values that range randomly with equal probability from 1 to 9. Therefore, the 90 customers have two patterns of preferences: "clustered" for products A, B and "random" for products C, D, E and F. On the 90 × 6 data matrix so defined, the TA was applied by computing first the MCA and successively, by calculating the *K*-means algorithm on the first two components identified by the MCA.

Figure 2 shows the biplot of categories of the 6 variables named A, B, C, D, E, F and followed by a number between 1 and 9 to distinguish categories. The total loss (5) is 7.39.

It can be clearly seen from the biplot that the most relevant categories are those of the two variables A and B together with other categories, e.g. F7, C7, E9, D1 from variables F, C, E and D. Thus, the clustered and the random patterns of the customers are assorted and not clearly distinguishable in the biplot. Furthermore, TA tends to

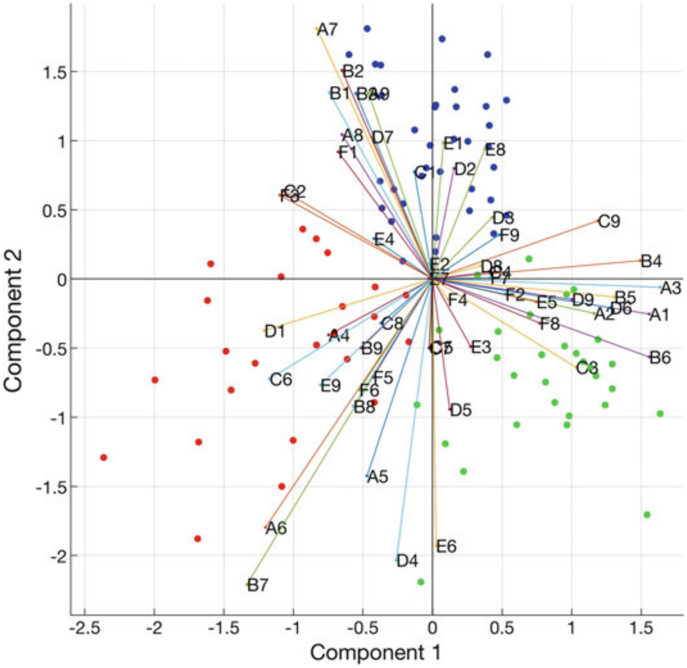


Fig. 2 Biplot of the 90 × 6 qualitative variables (A, B, C, D, E, F) with categories from 1 to 9. The three clusters are represented by three different colours

Table 1 Contingency table between *K*-Means groups and simulated groups

		K-Means			
		Group 1	Group 2	Group 3	Total
Simulated groups	Group 1	30	0	0	30
	Group 2	3	27	0	30
	Group 3	7	1	22	30
	Total	40	28	22	90

mask the three clusters of costumers, each one originally formed by 30 customers, as shown in Table 1.

In fact, the points classified in the three groups are 40, 28 and 22, respectively. Thus, 11 customers (12%) are misclassified (3 from the second cluster and 8 from the last cluster). The Adjusted Rand Index (ARI) between the generated three clusters and the three clusters obtained by *K*-means is $ARI = 0.6579$. Then, TA describes imprecisely the three clusters and defines components which do not clearly distinguish the two different preference patterns: the clustered, for products A, B and the random for the products C, D, E, F.

3 Multiple Correspondence *K*-Means model

Hwang et al. (2006) propose a convex combination the homogeneity criterion for MCA and the criterion for K-means; in this paper, let us use a different approach by specifying a model for the data, replacing Eq. (2) into Eq. (1). Thus, it follows that

$$J^{1/2}\mathbf{JBL}^{1/2} = (\mathbf{U}\bar{\mathbf{Y}} + \mathbf{E}_{KM})\mathbf{A}' + \mathbf{E}_{MCA}\ , \tag{6}$$

and rewriting the error term $\mathbf{E}_{MCKM} = \mathbf{E}_{KM}\mathbf{A}' + \mathbf{E}_{MCA}$, the resulting equation is here named multiple correspondence k-means (MCKM) model:

$$J^{1/2}\mathbf{JBL}^{1/2} = (\mathbf{U}\bar{\mathbf{Y}}\mathbf{A}' + \mathbf{E}_{MCKM})\ . \tag{7}$$

MCKM model identifies, simultaneously, the best partition of the N objects described by the best orthogonal linear combination of variables according to a single objective function. The coordinates of the projections onto the basis are given by the components y_{ip} collected in the matrix $\mathbf{Y} = \mathbf{XA}$. Within this subspace, hence, with these components, a partition of objects is sought such that the objects are "closest" to the centroids of the clusters (Vichi and Kiers 2001). When $\mathbf{X} = J^{1/2}\mathbf{JBL}^{1/2}$ is actually a quantitative data matrix, the least-squares (LS) estimation of model (7) is equal to the reduced k-means (RKM) model, proposed by De Soete and Carroll (1994). Additionally, when equation (7) is post-multiplied both sides by A, the RKM model is transformed into the factorial k-means (FKM) model, proposed by Vichi and Kiers (2001). Both models have been formalized for numeric data.

The LS estimation of MCKM corresponds to minimize the objective function

$$\begin{cases} ||J^{1/2}\mathbf{JBL}^{1/2} - \mathbf{U}\bar{\mathbf{Y}}A'||^2 \underset{\mathbf{A},\mathbf{U},\bar{\mathbf{Y}}}{\rightarrow} min \\ \mathbf{A}'\mathbf{A} = \mathbf{I}_P \\ \mathbf{U} \in \{0, 1\} \\ \mathbf{U}\mathbf{1}_K = \mathbf{1}_N \end{cases} . \tag{8}$$

4 Alternating Least-Squares Algorithm

The quadratic constrained problem of minimizing (8) can be solved by an alternative least-squares (ALS) algorithm, which is structured on three steps, as follows:

Step 0: Firstly, initial values are chosen for $\mathbf{A}$, $\mathbf{U}$ and $\bar{\mathbf{Y}}$; in particular, initial values for $\mathbf{A}$ and $\mathbf{U}$ can be chosen randomly satisfying the constraints shown in (8), while initial values are then given at once by $(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{Y}$.

Step 1: Minimize $F([u_{ik}]) = ||J^{1/2}\mathbf{JBL}^{1/2} - \mathbf{U}\bar{Y}A'||^2$ with respect to $\mathbf{U}$, given the current values of $\mathbf{A}$ and $\bar{\mathbf{Y}}$. The problem is solved for the rows of $\mathbf{U}$

independently by taking $u_{ik} = 1$ if $F([u_{ik}]) = min\{F([u_{iv}]) : v = 1, \dots, P; (v \neq k)\}$; $u_{ik} = 0$, otherwise.

Step 2: Given $\mathbf{U}$, update $\mathbf{A}$ and implicitly $\bar{\mathbf{Y}}$ by minimizing (8). The problem is solved by taking the first p eigenvectors of $\mathbf{X}'(\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}')\mathbf{X}$ (e.g. see Vichi and Kiers 2001).

Step 3: Compute the objective function (8) for the current values of $\mathbf{A}$, $\mathbf{U}$ and $\bar{\mathbf{Y}}$. When the updates of $\mathbf{A}$, $\mathbf{U}$ and $\bar{\mathbf{Y}}$ have decreased the function value, repeat the step 1 and 2; otherwise, the process has converged.

ALS algorithm monotonically decreases the loss function and, because the constraints on $\mathbf{U}$, the method can be expected to be rather sensitive to local optima. For these reasons, it is recommended the use of many randomly started runs to find the best solution. In some test, it has been valued that, for a good solution (a good local optimal value), the use of 500 random starts usually suffices.

5 Theoretical and Applied Properties

5.1 Theoretical Property

Property 1 *The LS solution of MCKM obtained by solving the quadratic problem (8) subject to constraints* $\mathbf{A}'\mathbf{A} = \mathbf{I}_P$, $\mathbf{U} \in \{0, 1\}$, *and* $\mathbf{U}\mathbf{1}_K = \mathbf{1}_N$ *is equivalent to the minimization of the objective function (5) used to give an overall estimation of the loss produced by Tandem Analysis results. In other terms, it can be proved the equality*

$$2f(\hat{\mathbf{Y}}, \hat{\mathbf{A}}, \hat{\mathbf{U}}, \hat{\bar{\mathbf{Y}}}) = ||J^{1/2}\mathbf{JBL}^{1/2} - \hat{\mathbf{Y}}\hat{\mathbf{A}}'||^2 + ||\hat{\mathbf{Y}} - \hat{\mathbf{U}}\hat{\bar{\mathbf{Y}}}||^2 = ||\mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}'||^2 \,, \tag{9}$$

where $\mathbf{X} = J^{1/2}\mathbf{JBL}^{1/2}$.

Proof In fact, after some algebra the objective function of MCKM can be written as

$$||\mathbf{X} - \mathbf{U}\bar{\mathbf{Y}}\mathbf{A}'||^2 = ||\mathbf{X} - \mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}'||^2 = tr(\mathbf{X}'\mathbf{X}) - tr(\mathbf{X}'\mathbf{U}\bar{\mathbf{X}}\mathbf{A}\mathbf{A}') \,. \tag{10}$$

Thus, it is necessary to prove that the objective function of the TA is equal to (10).

$$\begin{aligned} &||\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}'||^2 + ||\mathbf{X}\mathbf{A} - \mathbf{U}\bar{\mathbf{X}}\mathbf{A}||^2 = \\ &tr\{(\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}')'(\mathbf{X} - \mathbf{X}\mathbf{A}\mathbf{A}')\} + tr\{(\mathbf{X}\mathbf{A} - \mathbf{U}\bar{\mathbf{X}}\mathbf{A})'(\mathbf{X}\mathbf{A} - \mathbf{U}\bar{\mathbf{X}}\mathbf{A})\} = \\ &tr(\mathbf{X}'\mathbf{X}) - tr(\mathbf{X}'\mathbf{X}\mathbf{A}\mathbf{A}') - tr(\mathbf{A}\mathbf{A}'\mathbf{X}'\mathbf{X}) + tr(\mathbf{A}\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A}\mathbf{A}') + \\ &+ tr(\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A}) - tr(\mathbf{A}'\mathbf{X}'\mathbf{U}\bar{\mathbf{X}}\mathbf{A}) - tr(\mathbf{A}'\bar{\mathbf{X}}'\mathbf{U}'\mathbf{X}\mathbf{A}) + tr(\mathbf{A}'\bar{\mathbf{X}}'\mathbf{U}'\mathbf{U}\bar{\mathbf{X}}\mathbf{A}) \,. \end{aligned} \tag{11}$$

Now, knowing that $\mathbf{U}\bar{\mathbf{X}} = \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{X} = \mathbf{P_U X}$, where $\mathbf{P_U}$ the idempotent projector of matrix $\mathbf{U}$, equation (11) can be written as

$$\begin{aligned} & tr(\mathbf{X}'\mathbf{X}) - tr(\mathbf{A}'\mathbf{X}'\mathbf{XA}) - tr(\mathbf{AA}'\mathbf{X}'\mathbf{X}) + tr(\mathbf{A}'\mathbf{X}'\mathbf{XA}) + \\ & + tr(\mathbf{AA}'\mathbf{X}'\mathbf{X}) - tr(\mathbf{A}'\mathbf{X}'\mathbf{P_U XA}) - tr(\mathbf{A}'\mathbf{X}'\mathbf{P_U XA}) + tr(\mathbf{A}'\mathbf{X}'\mathbf{P_U P_U XA}) = \\ & = tr(\mathbf{X}'\mathbf{X}) - tr(\mathbf{A}'\mathbf{X}'\mathbf{P_U XA}) - tr(\mathbf{A}'\mathbf{X}'\mathbf{P_U XA}) + tr(\mathbf{A}'\mathbf{X}'\mathbf{P_U XA}) = \\ & = tr(\mathbf{X}'\mathbf{X}) - tr(\mathbf{X}'\mathbf{U}\bar{\mathbf{X}}\mathbf{AA}')\ , \end{aligned} \tag{12}$$

which complete the proof.

5.2 *Applied Property*

Let us apply the multiple correspondence *k*-means on the 90 × 6 data set used in Sect. 2 to show the limits of the Tandem Analysis in case categorical data are considered. The loss function (8) is equal to 7.23, better than the loss of the TA, with an improvement of the loss function of 2%. Even if the improvement seems small this time the biplot of multiple correspondence *k*-means in Fig. 3 shows a very clear synthesis of the data. Categories of products A and B are well distinguished from categories of products C, D, E, F, and therefore, the clustered and random patterns of preferences of customers are clearly differentiated. Furthermore, the clustering structure of the customers is well represented in the biplot. In fact, the three clusters are formed each one by 30 customers, as expected, and they are more homogeneous and well separated with respect to the clusters in the biplot of TA (Fig. 3).

The red cluster is formed by customers who like products A and are neutral on the product B (the first 30 rows, in the data set). The blue cluster is formed by customers who like the second product B and dislike the first product A (the second 30 rows of the data set). Finally, the green cluster of customers is formed by persons that dislike the product B and are neutral of on product A (the third and last 30 rows of the data set). So this time no misclassifications are observed for the clusters (see Table 2) and the two different patterns of products are differently represented in the plot as expected.

Table 2 Contingency table between MCKM groups and simulated groups

		K-Means			
		Group 1	Group 2	Group 3	Total
Simulated groups	Group 1	30	0	0	30
	Group 2	0	30	0	30
	Group 3	0	0	30	30
	Total	30	30	30	90

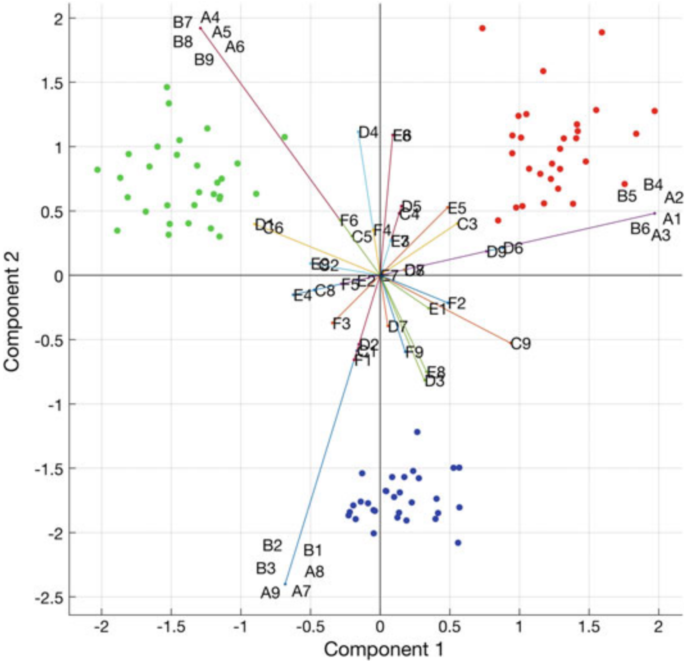


Fig. 3 Biplot of the multiple correspondence *K*-means . It can be clearly observed that the three clusters are homogeneous and well separated

6 Application on South Korean Underwear Manufacturer

The empirical data presented in this section is part of a large survey conducted by a South Korean underwear manufacturer in 1997 (Hwang et al. 2006), where 664 South Korean consumers were asked to provide responses for three multiple-choice items.

In particular, the first item asked which of eight brands of underwear the consumer most prefers (A): (A01) BYC, (A02) TRY, (A03) VICMAN, (A04) James Dean, (A05) Michiko-London, (A06) Benetton, (A07) Bodyguard and (A08) Calvin Klein; then, both domestic (A01, A02, A03, A04 and A07) and international (A05, A06 and A08) brands were included. The second item asked the attribute of underwear most sought by the consumers (B): (B01) comfortable, (B02) smooth, (B03) superior fabrics, (B04) reasonable price, (B05) fashionable design, (B06) favourable advertisements, (B07) trendy colour, (B08) good design, (B09) various colours, (B10) elastic, (B11) store is near, (B12) excellent fit, (B13) design quality, (B14) youth appeal and (B15) various sizes. The last item asked the age class of each consumer (C): (C01) 10–29, (C02) 30–49 and (C03) 50 and over. In Table 3, the frequency distributions of the three categorical variables are shown.

Table 3 Frequency distributions of the South Korean underwear manufacturer data

Brand (A)		Attributes (B)		Age (C)	
A01. BYC	201	B01. Comfortable	398	C01. 10–29	239
A02. TRY	131	B02. Smooth	65	C02. 30–49	242
A03. VICMAN	30	B03. Superior fabrics	29	C03. 50 and over	183
A04. James Dean	72	B04. Reasonable price	33		
A05. Michiko-London	11	B05. Fashionable design	67		
A06. Benetton	13	B06. Favourable advertisements	7		
A07. Bodyguard	166	B07. Trendy colour	15		
A08. Calvin Klein	40	B08. Good design	4		
		B09. Various colours	4		
		B10. Elastic	11		
		B11. Store is near	3		
		B12. Excellent fit	20		
		B13. Design quality	6		
		B14. Youth appeal	1		
		B15. Various sizes	1		

Table 4 Results of the MCA model applied on the South Korean underwear manufacturer data

Singular value	Inertia	Chi-square	Inertia (%)	Cum. inertia (%)
0.726	0.527	1048.930	6.870	6.870
0.644	0.414	824.878	5.400	12.270
Total	0.941	1873.808	12.270	–

P-value= 0 Degrees of freedom= 196

The analysis starts with the application of multiple correspondence analysis and, subsequently, the application of *K*-Means on the computed scores (Tandem Analysis). Hwang et al. (2006), suggested to apply MCA by fixing the number of components equal to 2 since sizes of the adjusted inertias appeared to decrease slowly after the first two. The results obtained by the MCA are shown in Table 4.

From Table 4, it is worthy to note that the non-revaluated explained variance of the two computed factors is equal to 12.27% of the total inertia (note that Greenacre 1984 recommends to adjust the inertias greater than 1/J using Benzecri 1979 formula). In Table 5, it is possible to observe the computed loadings among the two components and each category of the data.

From the table, it is easy to note that the categories with bigger contributions on the first component are the first two brands of underwear (A01 and A02) and the seventh brand (A07); the fifth attribute (B05); and the first and third class of the age (C01 and C03) whereas the categories with bigger contribution on the second

Table 5 Loading matrix of the MCA model applied on the South Korean underwear manufacturer data

Component 1			Component 2		
Brand	Attributes	Age	Brand	Attributes	Age
−0.250	−0.133	0.467	0.177	−0.152	0.102
−0.302	−0.065	−0.163	0.090	0.184	−0.374
−0.134	−0.008	−0.346	−0.363	0.285	0.312
0.135	−0.047	–	−0.291	0.234	–
0.161	0.373	–	0.311	0.064	–
0.181	−0.046	–	−0.031	−0.036	–
0.334	0.108	–	0.038	0.030	–
0.175	0.123	–	−0.077	0.017	–
–	−0.097	–	–	0.027	–
–	−0.082	–	–	−0.278	–
–	−0.020	–	–	0.162	–
–	−0.002	–	–	−0.164	–
–	0.152	–	–	−0.231	–
–	0.099	–	–	0.049	–
–	−0.067	–	–	−0.073	–

component are the third, fourth and fifth brand (A03, A04 and A05); the third, fourth, tenth and thirteenth attribute (B03, B04, B10 and B13); and second and third class of the age (C01 and C03). Then, the two component scores represent a very high number of the categories. However, the variables brands (A) and age (C) are more represented than attributes (B). Subsequently, according to the TA approach, the K-Means model on the two component scores has been applied. The fixed number of groups is $K = 3$ as suggested by Hwang et al. (2006).

The plot in Fig. 4 shows the projection of the single category on the bi-dimensional factorial plane and the distributions of the computed scores. We can note that the three defined groups are underlined with different colours.

The biplot shows that the groups are not well separated and they are characterized by an high inside heterogeneity. In fact, it is very hard to understand the preferences of the consumers that belong to the three groups.

Different results have been obtained with the multiple correspondence k-means approach. Fixing the same number of components and groups, the explained variance of the two components is around to 20%. The component loadings of the MCKM are represented in Table 6.

In the MCKM model, the categories with bigger contributions on the first component are the first two brands of underwear (A01 and A02) and the seventh brand (A07); the first and the third class of the age (C01 and C03). The categories with bigger contribution on the second component are the fourth, fifth, sixth, seventh and

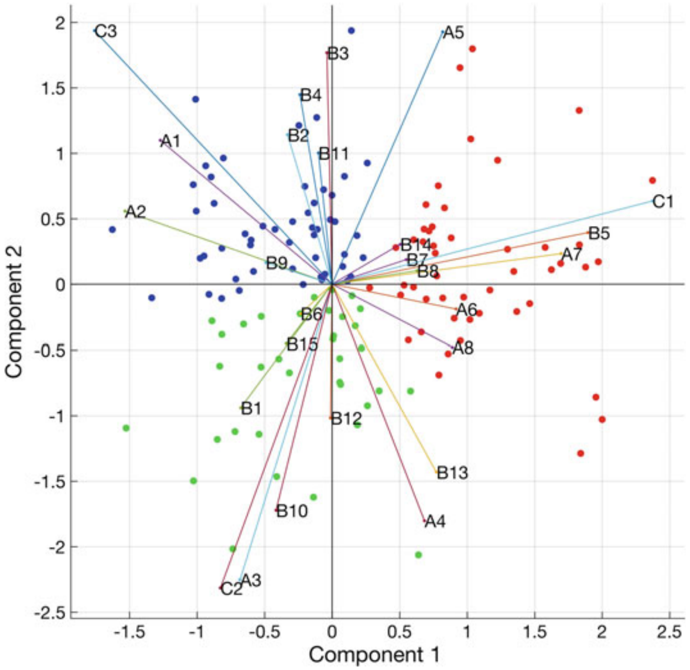


Fig. 4 Biplot of the sequential approach applied on South Korean underwear manufacturer data

Table 6 Loading matrix of the MCKM model applied on the South Korean underwear manufacturer data

Component 1			Component 2		
Brand	Attributes	Age	Brand	Attributes	Age
0.429	0.029	−0.252	0.159	0.040	−0.057
0.346	0.028	0.062	0.128	0.068	−0.018
0.158	0.034	0.216	0.046	−0.045	0.086
−0.123	0.025	–	−0.609	0.007	–
−0.048	−0.161	–	−0.238	−0.074	–
−0.052	0.031	–	−0.259	0.007	–
−0.694	−0.016	–	0.449	−0.018	–
−0.092	−0.046	–	−0.454	0.005	–
–	0.061	–	–	0.022	–
–	0.011	–	–	0.034	–
–	0.052	–	–	0.019	–
–	0.036	–	–	−0.093	–
–	−0.052	–	–	−0.132	–
–	−0.054	–	–	0.035	–
–	0.030	–	–	0.011	–

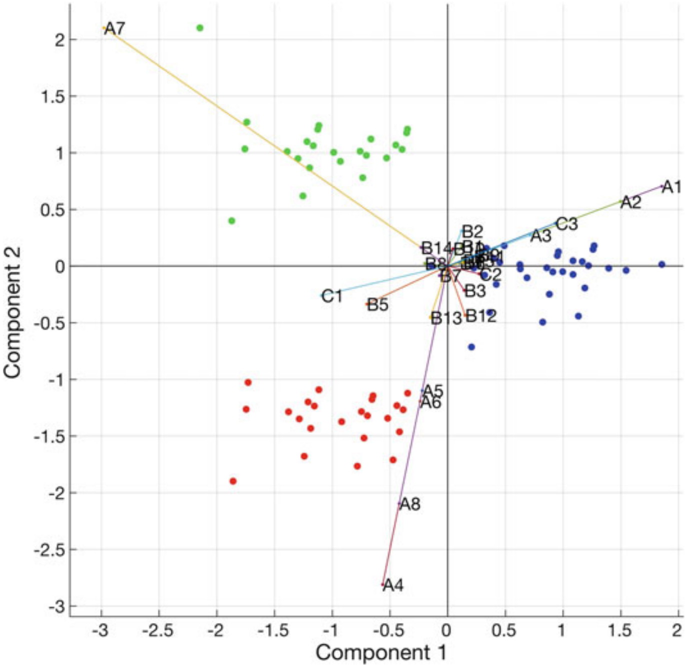


Fig. 5 Biplot of the simultaneous approach applied on South Korean underwear manufacturer data

eighth brand (A04, A05, A06, A07 and A8) only. Then, unlike TA, in the MCKM model the variable attributes (B) do not give a relevant contribution.

In Fig. 5 is shown the biplot where are represented the component scores and the three defined groups.

From the plot, we can note that the groups are well separated and homogeneous. In fact, it easy to note that the green group (166 observations) are the consumers that prefer the seventh brand (A07); the blue group (361 observations) are the consumers that prefer the first three brands (A01, A02 and A03) and they have mainly an age of 50 years and over (C03); finally, the red groups (137 observations) are the consumers that prefer the fourth, fifth, sixth and eight brand (A04, A05, A06, and A08). It is possible to verify these results observing the frequency distributions of the three categorical variables shown in Table 3.

7 Conclusions

Tandem Analysis (TA) is a well-known sequential procedure for clustering and dimensional reduction. It is frequently used in applications for quantitative data, however is has several limitations, because in some case it could identify factors that do not necessarily include all the information on the clustering structure of units. In particular, it can fail to find the correct clustering structure with a reduced set of factors (Vichi and Kiers 2001). TA is also frequently used when categorical variables are considered. It corresponds to apply MCA on the original data and successively K-means clustering on the component score matrix of MCA. In this paper, it was proved that also this TA has serious problems to correctly classify units and synthesize the relationships of the observed categorical variables. Thus, a model called multiple correspondence k-means (MCKM) was proposed and estimated in the LS by using an ALS algorithm. Property 1 proves that the LS estimation of MCKM corresponds to optimize the loss function of the TA which is only imprecisely estimated by the sequential application of MCA and K-means.

References

Arabie, P. (1994). Cluster analysis in marketing research. *Advanced Methods in Marketing Research*, 160–189.

Benzécri, J. P. (1979). Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, addendum et erratum à [BIN. MULT.]. *Les cahiers de l'analyse des données, 4*(3), 377–378.

Bolton, R. J., Krzanowski, W. J. (2012). Projection pursuit clustering for exploratory data analysis. *Journal of Computational and Graphical Statistics*.

Desarbo, W., Jedidi, K., Cool, K., & Schendel, D. (1991). Simultaneous multidimensional unfolding and cluster analysis: An investigation of strategic groups. *Marketing Letters*, *2*(2), 129–146.

De Soete, G., & Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. *New approaches in classification and data analysis* (pp. 212–219). Berlin, Heidelberg: Springer.

De Soete, G., & Heiser, W. J. (1993). A latent class unfolding model for analyzing single stimulus preference ratings. *Psychometrika*, *58*(4), 545–565.

Greenacre, M. J. (1984) *Theory and applications of correspondence analysis*.

Heiser, W. J. (1993). Clustering in low-dimensional space. *Information and classification* (pp. 162–173). Berlin, Heidelberg: Springer.

Hwang, H., Dillon, W. R., & Takane, Y. (2006). An extension of multiple correspondence analysis for identifying heterogeneous subgroups of respondents. *Psychometrika*, *71*(1), 161–171.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, (Vol. 1, no. 14, pp. 281–297).

Timmerman, M. E., Ceulemans, E., Kiers, H. A., & Vichi, M. (2010). Factorial and reduced K-means reconsidered. *Computational Statistics & Data Analysis*, *54*(7), 1858–1871.

Vichi, M., Kiers, H. A.: Factorial K-means analysis for two-way data. *Computational Statistics & Data Analysis*, **37(1)**, pp. 49–64 (2001). *Bioinformatics*, *17*(9), 763–774 (2001).

TaLTaC 3.0. A Multi-level Web Platform for Textual Big Data in the Social Sciences

Sergio Bolasco and Giovanni De Gasperis

Abstract The TaLTaC software package as a tool of lexical and textual analysis, versions 1.0 e 2.0, lived over the last decades (1999–2015). It appears now to have met its technological limits. The TaLTaC version 3.0 (from now on T3) has been redesigned to overcome those limits. The process included: (i) recoding of all inner software components with modern web-related languages and standards; (ii) adoption of a new kind of database (NoSQL) capable to handle corpora in the order of magnitude of gigabytes; (iii) new criteria for data storage and data processing. The software architecture is modular and allows to decouple user interaction from actual data computing. The two main components are: the GUI (graphical user interface), based on HTML5/CSS/Js and the back-end processing *CORE*. The new design also made it possible to run T3 among the mainstream operating systems: Os X, Windows, and Linux. From a single parsing operation, T3 produces many vocabularies for multi-level lexical analysis. This allows one to disambiguate, in a semiautomatic fashion, between the different text graphical forms on the basis of concordance. I also allows for a virtual transformation of simple forms into multi-words.

Keywords Software engineering · Text mining · Big data · Lexical analysis

1 Introduction

In a near future, the perspective of text mining is to work with so-called big data. Currently, files of sizes in the order of gigabyte are available, containing corpora of millions of texts. These corpora constitute vocabularies of millions of single types,or

S. Bolasco (✉)
Dipartimento MEMOTEF, Università degli Studi di Roma “La Sapienza”, Rome, Italy
e-mail: sergio.bolasco@uniroma1.it

G. De Gasperis
Dipartimento di Ingegneria e Scienze dell’ Informazione e Matematica,
Università degli Studi dell’ Aquila, L’ Aquila, Italy
e-mail: giovanni.degasperis@univaq.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_9

billions of tokens, or even entire repositories of terabytes of data, containing billions of records (Mayer-Schönberger and Cukier 2013).

Since the beginning of the statistical analysis of textual data (Lebart and Salem 1988), the vocabularies of a corpus constituted of “micro” big data (tens/hundreds of thousands of types deriving from files whose size was around a few megabytes): a vocabulary being a lexical statistical variable with tens of thousands of modalities. In the context of the analysis of textual data, the issue of being capable to display huge lists (vocabulary or concordance), or to extract from these vocabularies subsets of keywords (Bolasco 2013), has been always present.

Given these large numbers, quantitative measurements are extremely stable. In this context, the real problem is to reduce the amount of information only extracting useful information. This leads: (i) to work on the concepts or classes of words; (ii) to model variable structures (local grammars) (Silberztein 1993); (iii) to look at documents that are good examples of the relevant phenomena (clustering and categorization of texts); (iv) to also study “rare” phenomena through the extraction of concordance, which is sufficiently consistent—as experimentally shown in (Escoubas-Benveniste et al. 2012). In this sense, it is possible to produce effective text mining solutions, by matching unstructured information with the structured one. In order to achieve this objective, the ETL processes are critical.

2 The Design and Evolution of the TaLTaC Project

The T3 project is the evolution of the TaLTaC2 package developed between 1999 and 2015 (Bolasco 2010, 2013). It is based on the multi-tier software engineering practices described in the work of Allier et al. (2015), which emphasize modularity, reusability, and usability typical of web-based applications. In this way, highly complex applications can be reduced to simpler modules, interconnected at multi-tier stages through high-performance communication channels. Even though this may seem not appropriate for a desktop application, it is mostly suited for an application such as TaLTaC, which shall evolve into many possible deployment configurations: from the desktop context to the shared configuration of a typical computer room in a data science laboratory up to the online applications-as-a-service solution (Bolasco et al. 2016). For this reasons, we adopted the following design guide lines for the evolution of the TaLTaC package: (i) to be a multi-platform software; (ii) to have theoretically no limits of corpora size to be analyzed (big data); (iii) to maximize the exploitation of the computational capabilities of the underlying hardware; (iv) to be an inherently distributed application, so as to allow different deployment configurations. The improved modularity shall allow to introduce layered notations and virtual lexicalization in the type vocabulary, which we call “lexical multi-level analysis.”

2.1 Software Engineering of TaLTaC3

Under the main requirements described above, we introduced an agile software engineering methodology, building up over the "TaLTaC user experience" (Bolasco 2010): (i) the GUI design is inspired by the TaLTaC2 "look and feel," while increasing the usability in the human–computer interaction, through the adoption of recent web-based techniques; (ii) TaLTaC2 text analysis algorithms have been ported to web-compatible frameworks and languages of T3; (iii) T3 includes scalability technologies as NoSQL data storage (which will be described later in this section), thus improving performance. The software technologies and languages adopted in T3 include:

- *HTML5* standard for the GUI;
- *jQuery* and its derived Javascript frameworks to encapsulate the GUI user interaction functions;
- *JSON (JavaScript Object Notation)*: as an inter-module language standard, with a structured and agile format for data exchange in client/server applications;
- *Python/PyPy*: advanced script/compiled programming language, mostly used for textual data analysis and natural language processing at the CORE back end;
- *NOSQL*: high-performance key/value data structure server adopted for vocabularies/linguistic resources persistence;
- *RESTful*: interface standard for data exchange over the HTTP web protocol;
- *Multi-processing*: exploiting in the best possible way multi-core hardware, which can finally be considered "off-the-shelf" technology even in personal computers.

3 TaLTaC3 General Software Architecture

As described earlier, T3 has been designed from the ground up to be extremely modular in order to fulfill the desired requirements and design specifications. The software architecture is shown in Fig. 1. On the left, the GUI is represented by an Internet browser content window, and, on the right, the CORE is composed of several running processes, each dedicated to a specific task/function (tokenization, normalization, vocabulary extraction, alignment of lexical multilayers, storage management, and data processing algorithms library).

4 TaLTaC3 Innovation with Respect to TaLTaC2

The proposed software architecture facilitated the implementation of the different specifications. Firstly, the *UTF-8* characters coding allows multi-language analysis. Secondly, a greater integration of functions is obtainable due to an object-oriented

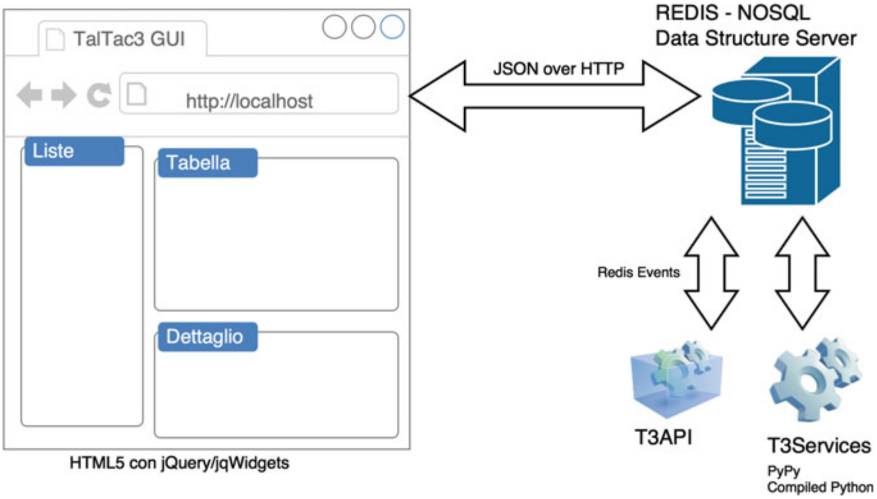


Fig. 1 TaLTaC3 software architecture

paradigm. From a single parse, T3 produces many vocabularies for multi-level lexical analysis, enabling the possibility to work independently at different layers. More specifically, the corpus tokenization generates the base vocabulary representing the documental study level (each type is the original graphic text). The next level, named layer L0, provides the "pre-normalized" vocabulary optimizing original text in words, like final words with apostrophes in words accented, upper case to lower case, number reconstruction in a single token (for example: 3.15; −2.7%; 8,456.00; 1,500,000). The level below, named layer L1, produces the understood vocabulary, transforming the words where necessary in units of meaning (named entities, multiword, locutions). In this way, the lexicalization (the reconstruction of an expression in a single occurrence) is virtual and not "hard coded," augmenting the representation structures of the original types after processing. Furthermore, one can implement additional lexical levels. The layer L2 produces linguistic annotations such as the t-uple {type, PoS, headword}. In this way, T3 allows to exactly disambiguate, in a semiautomatic fashion, specific occurrences on the basis of concordance. Analogously, the layer L3 contains semantic annotations such as {type, theme vs topic vs synset}, useful for creating word classes that are thematically homogeneous.

4.1 TaLTaC3.0 Main Screen

The main screen of T3 is divided into a main content area to the right and a left side bar to access textual resources and data. Through the side bar, it is possible to select the job session and/or resources (linguistic or statistical meta-data for comparisons in order to select sets of keywords); and below there are the vocabulary layers on the bottom area, including information on the current state of the session and some lexicometric measures are shown. On the right side, the main area of the detailed output of the running computing process is shown. For example, in the main area a browsable vocabulary, where a single type can be selected, and its concordance can be shown inside the bottom area.

5 TaLTaC3.0 Benchmark

The benchmark of T3 includes several tests aimed to show CPU time and memory used, with respect to get the corpus vocabulary of various sizes on the equivalent hardware platform, but different OS configurations, also comparing it to the predecessor TaLTaC2 (from now on T2). The tests involve the first tokenization that produces only the basic vocabulary by T2 and the vocabularies of three layers (base + L0 + L1) by T3.

A small Italian corpus of 100 newspaper articles from La Repubblica made of 0.4 MB file and 70,000 tokens is processed by T2 in 16 s, while T3 in the same machine requires only 0.5 s, equivalent to 32× computational efficiency improvement. The corpus which collects the 72 chapters of the Bible (a file of about 5 MB and 883,900 tokens) is processed by T2 in 24 s, and by T3 in 2.2 s, with 10× improvement. Finally, a corpus, collected by ISTAT (Italian National Institute of Statistics), of 240 thousands of tweets, made of 30 MB file and 6 millions of tokens, was computed by T3 in 41 s, while it cannot be computed by T2 since the latter's limitation to 100 thousands of context units.

Indeed, computation times do not constitute an issue for middle-sized corpora, exploiting the new design and information technologies which also takes advantages of multi-core systems and cluster of computers for big data. For example, the TeraStat [1] consists of eighteen cluster compute nodes, for a computational capacity of over 5 Teraflops/s and a total capacity of disk space of over 80 terabytes. On TeraStat, T3 will elaborate corpus of several gigabytes in absolutely reasonable times. Even now, without taking advantage of the multiple cores, a corpus of half a billion words (a file of about 3 Giga Bytes) is processed by T3 on a MAC Pro (Mac OS X 10.5, Intel Core i7, 2.7 GHz, 16 GB RAM, SSD 512 GB) in 25 min.

5.1 Some Results on Comparable Hardware Architecture in Function of Corpus Size

In two corpora of newspaper articles (40 MB each file), T2 calculation times depend on the document size. To compute the first corpus composed of 100 large documents

[1] http://www.dss.uniroma1.it/en/node/6365 (2016). Accessed 30 Apr 2016.

(groups of articles) of the French newspaper “Le Monde,” T2 requires a 45% longer CPU time than the second corpus with equal total amount of occurrences but made of 10,000 small documents (single articles) of the Italian newspaper “La Repubblica.” On the contrary, with T3, the first corpus requires less processing time than the latter. Therefore, fewer documents—regardless of their size—require less time to tokenize, with the same number of total occurrences. We think this is a result of the new architecture of the T3 software. This property, as you can see in Table 1, is not homogeneously verified over different operating systems. While Windows just doubles the computing time, Linux has a constant performance and OS X shows overall a better efficiency and does not double the CPU time.

The tests so far show a significant linearity of computing times, in function of number of documents (“La Repubblica” articles) and corpus size. Computing time linearity is as follows: from 0.155 GB (1 year of “La Repubblica”: 1992) to 1.51 GB (10 years of “La Repubblica”: 1990–1999), respectively, from 1 min 18 s (1 year) to 14 min 44 s (10 years). See details in Table 2.

Table 1 OS performance comparison (in s)

Corpus	Size	Windows[a] (s)	Linux[b] (s)	Os X[c] (s)
Le Monde	100 documents	17.3	25.1	16.3
La Repubblica	10’000 documents	35.0	25.4	21.6

[a] Windows 7, Intel Core i7, 2.6 GHz, 16 GB RAM, SSD 512 GB (Toshiba Z30)
[b] Linux Mint MATE 64 8 (Rose), Intel Core Xeon, 2.93 GHz, 24 GB RAM, HD 1TB (Dell 5500)
[c] Mac Os X 10.5, Intel Core i/, 2.7 GHz, 16 GB RAM, SSD 512 GB (Mac Book Pro)

Table 2 Input size performance comparison (in s)

Corpus	Size	Windows[a]	Linux[b] (s)	Os X[c] (s)	Tokens[d] (M)
La Republica	37’000 articles	135.2 s	93.9	78.2	28.8
La Republica	100’000 articles	270.4 s	242.2	202.1	74.1
La Republica	150’000 articles	406.0 s	355.2	284.0	106.4
La Republica	400’000 articles	n.a.	986.0	844.0	278.9

[a] Windows 7, Intel Core i7, 2.6 GHz, 16 GB RAM, SSD 512 GB (Toshiba Z30)
[b] Linux Mint MATE 64 8 (Rose), Intel Core Xeon, 2.93 GHz, 24 GB RAM, HD 1TB (Dell 5500)
[c] Mac Os X 10.5, Intel Core i/, 2.7 GHz, 16 GB RAM, SSD 512 GB (Mac Book Pro)
[d] In millions

6 Conclusions

By observing the first performance data of T3, we can conclude that: (a) the adopted key/value data persistent subsystem is robust, also in the case when the user needs to use to different sessions at the same time; (b) computing times are already satisfactory just using a single core during parsing and two cores during lexical normalization, laying the base for better performances when a full multi-core optimization will be completed, mostly suited for big data application.

The multi-layer logic allows to extend the computational capability over a wider range of research activity, increasing scientific accuracy of textual data comparison and computation results between several analysis phases. For example, at layer L3 corresponding to semantic tagging, it is possible to directly calculate specificity over "concepts" (coherent topic lemmas) in respect of several partitions, without loosing information about specificity previously calculated on each type in layers L0 and L1 or on verbs lemmas into layer L2.

Also, T3 has been conveniently packaged as an universal desktop application adopting the Electron [2] multi-platform approach of a browser-based application container; it takes care of starting all the necessary back-end processes when running in a desktop environment, independently of the underlining operating system.

References

Allier, S., Barais, O., Baudry, B., Bourcier, J., Daubert, E., Fleurey, F., et al. (2015). Multitier Diversification in Web-Based Software Applications. *IEEE Software*, *32*(1), 83–90.

Bolasco, S. (2010). Taltac 2.10 Sviluppi, esperienze ed elementi essenziali di analisi automatica dei testi. Milano: Led.

Bolasco, S., Baiocchi, F., Canzonetti, A., & De Gasperis, G. (2016). TaLTaC 3.0: Un software multi-lessicale e uni-testuale ad architettura web. In D. Mayaffre, C. Poudat, L. Vanni, V. Magri & P. Follette (Eds.), Proceedings of 13th International Conference on Statistical Analysis of Textual Data, University Nice Sophia Antipolis, 225–235.

Bolasco, S. (2013). *L'analisi automatica dei testi. Fare ricerca con il text mining*. Roma: Carocci.

Escoubas-Benveniste, M. P., Floquet, O., & Bolasco, S. (2012). Contribution empirique à l'étude du gérondif et du participe présent en français parlé et écrit. JADT 2012: 11èmes Journées internationales d'Analyse statistique des Données Textuelles, 473–485.

Lebart, L., & Salem, A. (1988). *Analyse statistique des donnes textuelles*. Paris: Dunod.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data A Revolution that will transform how we live, work and think*. London: John Murray.

Silberztein, M. (1993). *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*. Paris: Masson.

[2] http://github.com/electron/electron, accessed November 2016.

Sparsity Data Reduction in Textual Network Analysis

An Exercise on Sustainability Meaning

Emma Zavarrone, Filomena Grassia, Maria Gabriella Grassia and Marina Marino

Abstract In this paper, we propose a new strategy to derive an unweighted adjacency matrix from an affiliation matrix. The strategy is based on the use of a biclustering technique in order to reduce the sparsity of the matrix without changing the network structure. As an example, we implemented this approach to seek the common meaning of the term sustainability by using an affiliation matrix characterized by a core–periphery structure. The application of BiMax biclustering algorithm shows a sparsity reduction of the unweighted adjacency matrix with an invariant network structure.

Keywords Network text analysis · Sparse matrices · Biclustering · Sustainability

1 Introduction

The term *network text analysis* (NTA) was employed from Diesner and Carley (2005) to describe a wide variety of *computer-supported solutions* that enable analysts to *extract networks of concepts* from texts and to discern the *meaning* represented or encoded therein. The key underlying assumption of such methods or solutions, they assert, is that the *language and knowledge* embodied in a text may be *modeled* as

E. Zavarrone (✉)
Department of Marketing, Communication and Consumer Behavior, University IULM, via Carlo Bo, Milan, Italy
e-mail: emma.zavarrone@iulm.it

F. Grassia
Italian National Institute of Statistics (ISTAT), Via C. Balbo, 00184 Rome, Italy
e-mail: grassia@istat.it

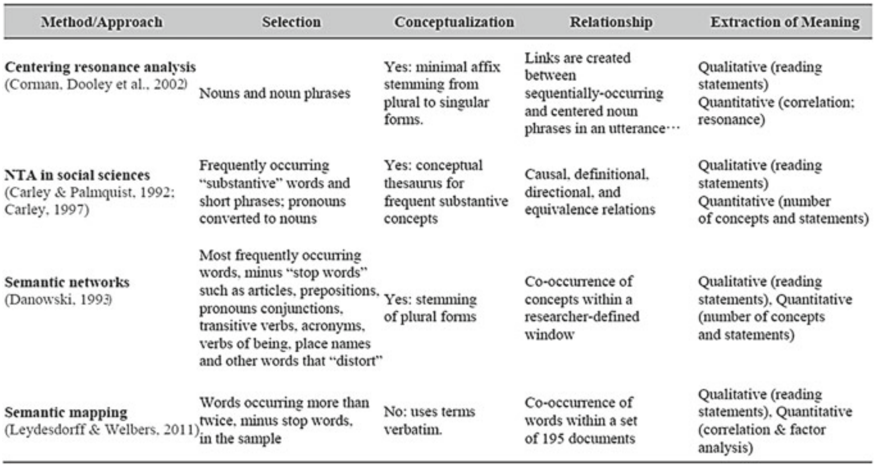
M.G. Grassia · M. Marino
Department of Social Science, University "Federico II", Vico Monte della Pietá, 80138 Naples, Italy
e-mail: mgrassia@unina.it

M. Marino
e-mail: mari@unina.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_10

a network *of words and the relations between them*. A second important assumption is that the position of concepts within a text network provides insight into the meaning or prominent themes of the text as a whole. Approaches to NTA differ with regard to how the steps of assigning of words and phrases to conceptual categories and assigning of links to pairs of concepts are performed. Other differences could be the level of automation or computer support, the linguistic unit of analysis, and the degree and basis of concept generalization (Hunter 2014). Concept, relationship, statement, and map (the four building blocks of text networks) are the same. Specifically, a concept is *an ideational kernel* often represented by *a single word or phrase*, while a relationship is *a tie that links two concepts together*. A statement is *two concepts and the relationship between them*, while a map is simply a *network formed from statements* (Carley and Palmquist 1992). There are many groups of methods for constructing networks of word. The most important are as follows: *Concept Maps* (Novak and Gowin 1984); *Hypertext* (Trigg and Weiser 1986); *Quantitative Narrative Analysis* (Franzosi 1989); *Network text analysis in the social sciences* (Carley and Palmquist 1992; Carley 1997); *Semantic Networks* (Danowski 1993); *Semantic Grammars* (Roberts 1997); *Semantic Webs* (Berners-Lee et al. 2001; *Centering Resonance Analysis* (Corman et al. 2002); *Network Analysis of Texts* (Batagelj and Mrvar 2007); *Automatic Quantitative Analysis* (Franzosi 2010; Sudhahar et al. 2011); *Network Textual Analysis* (Paranyushkin 2011); *Semantic Mapping* (Leydesdorff and Welbers 2011).

Hunter (2014), by these groups of methods, chose four (Centering Resonance Analysis (CRA), network text analysis in the social sciences, Semantic Networks, and Semantic Mapping) and summarized them along the following dimensions (Fig. 1):



Method/Approach	Selection	Conceptualization	Relationship	Extraction of Meaning
Centering resonance analysis (Corman, Dooley et al., 2002)	Nouns and noun phrases	Yes: minimal affix stemming from plural to singular forms.	Links are created between sequentially-occurring and centered noun phrases in an utterance···	Qualitative (reading statements) Quantitative (correlation; resonance)
NTA in social sciences (Carley & Palmquist, 1992; Carley, 1997)	Frequently occurring "substantive" words and short phrases; pronouns converted to nouns	Yes: conceptual thesaurus for frequent substantive concepts	Causal, definitional, directional, and equivalence relations	Qualitative (reading statements) Quantitative (number of concepts and statements)
Semantic networks (Danowski, 1993)	Most frequently occurring words, minus "stop words" such as articles, prepositions, pronouns conjunctions, transitive verbs, acronyms, verbs of being, place names and other words that "distort"	Yes: stemming of plural forms	Co-occurrence of concepts within a researcher-defined window	Qualitative (reading statements), Quantitative (number of concepts and statements)
Semantic mapping (Leydesdorff & Welbers, 2011)	Words occurring more than twice, minus stop words, in the sample	No: uses terms verbatim.	Co-occurrence of words within a set of 195 documents	Qualitative (reading statements), Quantitative (correlation & factor analysis)

Fig. 1 Selected methods for generating and analyzing text networks (Hunter 2014)

- *selection*, i.e., the basis upon which words are selected for or excluded from the subsequent analysis;
- *conceptualization*, i.e., to which conceptual categories are the selected words mapped;
- *relationships*, i.e., the basis upon which pairs of concepts are linked to one another;
- *extraction of meaning*.

Hunter said that although the sample was relatively small, it is possible to generalize broadly about these approaches to NTA. Firstly, all methods select some words for further analysis while excluding others from further consideration. Articles, prepositions, and pronouns are among the words most often excluded. The remaining words then are aggregated to higher-order conceptual categories, typically grammatical or logical in nature. Next, pairs of concepts are then related to one another, most frequently on the basis of their co-occurrence within some window of words. Finally, an analysis of the structure or pattern of these relationships through network analysis and/or statistical methods like factor analysis is employed to extract meaning. Like other forms of content analysis, NTA explicitly assumes that structure encodes meaning. Where NTA differs from traditional content analytic approaches, i.e., those concerned with word frequency, is that meaning is encoded in the structure of the network. Most specifically, in the extraction phase of NTA, prime importance is placed upon the position or role of concepts within the text network. In short, the more influential the network roles and position occupied by concepts, the greater the assumed thematic or semantic relevance they are assumed to have.

After the selection and conceptualization stages, for the *extraction of meaning*, in many of these methods, a *terms–documents* matrix (Zha and Zhang 2000) is built. This matrix (called *lexical table*) can be assimilated to an *affiliation matrix* in which the actors (rows) are the concepts, while the events (columns) represent the documents. So, the starting point of the analysis is a lexical table **T** $(p \times m)$ composed by p terms (rows), m documents (columns), and $p \times m$ cells that indicate the occurrences of the p terms in m parts of a corpus. There are two distinct approaches to analyze the lexical table. The first approach, called the conversion method, projects the two-mode network into two one-mode networks: The matrix **T** $(p \times m)$ becomes **W** $(p \times p)$ and **V** $(m \times m)$ with W = TT' and V = T'T. The projected datasets are analyzed using standard single-mode techniques. The second approach, called the direct method, analyzes the network directly with the two modes considered jointly. The direct approach in recent years has been preferred. However, Everett and Borgatti (2013) show that the conversion method is generally safe to use and often has conceptual advantages over the direct method.

Very often, when the documents are tweets or blog contents, and the language is Italian, the **T** matrix presents a high number of zero. The complex syntactic structures, the richness of the terms, and the resulting synonyms make the selection and conceptualization stages difficult, and the concepts are very numerous. A **T** matrix with a high number of zero values might distort the network, and consequently, the metrics might be biased. In other words, a T matrix should be characterized by a low level of sparsity. If it is not, a sparsity reduction should be done.

In this paper, we propose a strategy for reducing the sparsity of an affiliation matrix, preserving the textual network structure. To do it, we use a *biclustering* method.

The paper is organized as follows. In Sect. 2, we present a brief introduction to biclustering algorithms and heat map algorithms. In Sect. 3, we describe in detail our approach to dimensionality reduction. In Sect. 4, our strategy is implemented into an exercise focused on the concept of sustainable development. Finally, in Sect. 5, we discuss the proposal and provide suggestions for further developments of the research.

2 Biclustering and Heat Map Algorithms

A simultaneous clustering of rows and columns of two-dimensional data matrix is called *biclustering*. Names such as *direct clustering, co-clustering, two-mode clustering*, among others, are often used in the literature to refer to the same problem formulation.

One of the earliest biclustering formulations is the direct clustering algorithm introduced by Hartigan (1974) who introduced the principle to split the data matrix into submatrices and use the variance for assessing the quality of each partition.

A bicluster is a subset of rows characterized by similar behavior across a subset of columns, and *vice versa* (Madeira and Oliveira 2004).

Given an affiliation matrices matrix T with p rows and m columns, a bicluster A_{ij} is a subset of rows and a subset of columns of T:

$$A_{ij} = (I, J) \tag{1}$$

$$(I \subseteq p); (k \leq p) \tag{2}$$

$$(J \subseteq m); (s \leq m) \tag{3}$$

Biclustering allows identification of submatrices. Each submatrix is a subgroup of terms and subgroup of a documents, where the terms express highly correlated activities for every document. In recent years, several biclustering algorithms have been proposed. Madeira and Oliveira (2004) proposed a following classification of biclustering algorithms and relative end rules based on variance:

- Class1: Biclusters with constant values;
- Class2: Biclusters with constant values on rows or columns;
- Class3: Biclusters with coherent values; and
- Class4: Biclusters with coherent evolutions.

We use, for our sparsity reduction strategy, an algorithm proposed by Prelic et al. (2006), called BiMax. This algorithm belongs to the first category (biclusters with constant values). The objective of BiMax is intentionally simple: It finds all biclusters consisting entirely of 1 s in a binary matrix. Specifically, BiMax enumerates all inclusion-maximal biclusters, which are biclusters of all 1 s to which no row or column can be added without introducing 0 s. BiMax only works with binary matrices.

The most widespread visualization of the biclustering is heat map.

The biclustering heat map or heat map clustering can be defined as tiling of a data matrix with cluster trees appended to its margin (Wilkinson and Friendly 2009).

It is also used often for visualizing a term–document, a textual network. The designed area and the associated color allow us to classify in a fast way the correspondence between structure of term–document matrix and network. So a textual network could be investigated in a explorative perspective using heat maps since there is some correspondence between form of heat map and network structures.

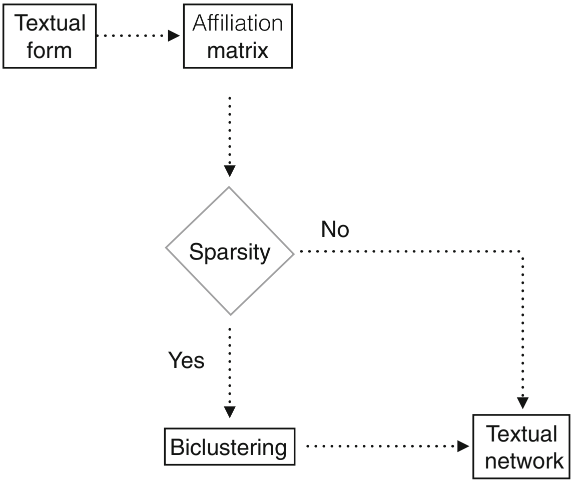
3 Our Proposal

Our methodological proposal starts from the **T** matrix, developing an ad hoc strategy to extract significant information from it. We trasform the **T** matrix in a binary matrix $\mathbf{T}_{B1}$ (0 if the occurrence is 0; 1 otherwise). We reduce the sparsity of the $\mathbf{T}_B$ matrix with the BiMax biclustering algorithm and then translate the relationship between terms and terms in a *one-mode unweighted* matrix $\mathbf{W}_B$ (Fig. 2).

The steps of our strategy are as follows:

1. *Selection* and *conceptualization* of the words. The text is preprocessed to transform the words in *textual forms* (Bolasco 1999):

Fig. 2 Flowchart of proposed methodology



- cleaning of the text (definition of alphabet characters/separators);
- normalizing the text (recognition of particular entities such as dates, acronyms, abbreviations); and
- introducing text annotation (introduction of meta-information by grammatical and semantic tagging, lemmatization, etc.).

2. Building the *terms–documents* matrix $\mathbf{T}$ $(p \times m)$.
3. Building the *binary terms–documents* matrix $\mathbf{T}_B$ $(p \times m)$. The $\mathbf{T}$ matrix was dichotomized. Applying the direct method, we can plot and analyze the *two-mode unweighted* N_{T_B} based on matrix $\mathbf{T}_B$.
4. Defining the *one-mode unweighted* matrix $\mathbf{W}$ $(p \times p)$ from $\mathbf{T}_B$ through a normalization quantile method (conversion approach).
5. Building the heat map of the $\mathbf{W}$ matrix to visualize its sparsity.
6. Plotting and analyzing the network N_W based on $\mathbf{W}$ matrix. In order to study both the cohesion of network and the links among terms, a basket of measures is selected (Freeman 1979):

   a. density;
   b. degree centrality;
   c. closeness centrality;
   d. standardized betweeness centrality; and
   e. cliques analysis.

7. Building the $\mathbf{T}_{B1}$ $(\mathrm{p}_1 \times \mathrm{m}_1)$ matrix. This matrix is obtained from the $\mathbf{T}$ $(p \times m)$ matrix with the BiMax algorithm (Prelic et al. 2006), where $p_1 \ll p$ and $m_1 \ll m$ indicate the presence of p_1 *textualform* (i.e., the concepts) in m_1 *conceptual* documents. In order to choose the most representative bipartition of terms and documents, $m(m-1)$ combinations of biclustering must be tested. The best combination with $p_1 \ll p$ and $m_1 \ll m$ will be chosen according to the highest value of *Jaccard* index.
8. Building the $\mathbf{W}_1$ $(\mathrm{p}_1 \times \mathrm{p}_1)$ matrix through a normalization quantile method. This matrix is a one-mode unweighted matrix, and it represents the relational system of the selected *textualform* (i.e., the biclustering selected concepts).
9. Building the heat map of the $\mathbf{W}_1$ matrix to visualize its sparsity.
10. Analyzing the network N_{W_1} based on $\mathbf{W}_1$ matrix.
11. Comparing the results for the network N_W and the network N_{W_1}.

The proposed strategy allows to reduce the sparsity of the term–document matrix with following results:

- simplifying the network structure;
- outlining a keyword basket; and
- identifying a cohesive subgroup of keywords.

4 Exercise on Sustainability Meaning

4.1 Sustainability Meaning

The institutional definition of *sustainable development* was first given by The World Conservation Strategy published in 1980 (International Union for Conservation of Nature and Natural Resources—IUCN, United Nations Environment Programme—UNEP, and World Wide Fund for Nature—WWF (1980)) and later reinforced by the World Commission on Environment and Development report (also known as the Brundtland Commission) that holds the key statement of sustainable development, defined as "*development that meets the needs of the present without compromising the ability of future generations to meet their own needs*" (Brundtland 1989). Over the years, the use of the term *sustainability* has been widely spread over several contexts (i.e., human rights, politics, and health) with a diverse spectrum of specific definition, even if several researchers (Pearce 1989; Middleton et al. 1993; Wackernagel and Rees 1996; Giddings et al. 2002) agree on the intentional vagueness of the WCED definition that was chosen precisely because it was ambiguous and thus accessible to a wide range of interest groups in society and in different countries. As suggested by Tilbury et al. (2002), it is open to widely different interpretation and its vagueness allows the attribution of many meanings revolving around questions such as "*Over what time period are we talking sustainability? The human life span? This generation and the next? Or are we concerned with sustainability on ecological timescales? And what kind of development do we want to sustain: social, cultural, political, spiritual, and/or economic? And are these separable? What changes are required to achieve sustainability and how are they to be achieved? What are the implications for economic growth? Are there limits to economic growth in a sustainable society and, if so, what are they?*"

However, the definition is still incomplete because it does not address:

- the environmental and ecological consequences (Murphy and Price 2005; Sumudo 2002);
- the relationship of environmental crises with environmental ethics and values in anthropocentric perspective (Buchdahl and Raper 1998; Sarvestani and Shahvali 2008; Seghezzo 2009; Vucetich and Nelson 2010);
- the sustainable development can be hidden if economic development is proposed from an egocentric point of view (Imran and Alam 2011); and
- the interrelationship between social, economic, and environmental contexts (Davidson 2011).

The vagueness of the definition of sustainable development has led to a complex framework of interpretations that does not capture the whole picture. This vagueness has also effects on the peoples' daily lives, since it often produces an (ab)use of the term sustainability in politics, ethics, economics, labor, communications, and other related fields.

4.2 The Data

Human rights, politics, and health are just few of the contexts in which the idea of sustainability is raised, so looking for the true or latent meanings of the term, as commonly used, is believed to be the first step toward an understanding of the concept. Our analysis investigates, through the blog contents, the common sense of the term *sustainability* in Italian language. The threats collected from 10 Italian blogs in which the term *sustainability* appeared represent the starting dataset. The survey was carried out in April 2015. Blogs were not monothematic and were not related to environmental matters. Overall, 40 treats were collected, and due to their brevity, only 57% was used for the analysis.

4.3 The Step of Our Strategy

Step One
We choose to put together threats coming from the same blog so we started from a *corpus* composed by ten blogs (i.e., documents). The text was preprocessed to transform the words in *textual forms* (Feinerer and Hornik 2014).

Step Two
We built the $\mathbf{T}$ matrix with $\mathbf{p} = 417$ *terms* and $\mathbf{m} = 10$ *documents*.

Step Three
We dichotomized the $\mathbf{T}$ matrix, and we built the *binary terms–documents* matrix $\mathbf{T}_B$. This matrix has a lot of zeroes with a sparsity level of 84%. The two-mode network visualization of $\mathbf{T}_B$ (Fig. 3) confirms the difficulty for a simple interpretation of the links among the terms and, at the same time, shows a core–periphery structure.
The analysis on core–periphery structure (Borgatti and Everett 1999) was made through the algorithm CORR in UCINET Borgatti et al. (2002). The correlation coefficient is low (0.32), probably, for the sparsity of $\mathbf{T}$.

Step Four
We defined the *one-mode unweighted* matrix $\mathbf{W}$ (417×417) through a normalization quantile method from $\mathbf{T}_B$.

Step Five
We built the heat map of the $\mathbf{W}$ matrix to visualize its sparsity (Fig. 4, *left side*).

Step Six
We analyzed the network N_W based on $\mathbf{W}$ matrix (Table 1, Figs. 5 and 6, *left side*). We used the *SNA* package Butts (2014).

Step Seven
We reduced the $\mathbf{T}_B$ (417×10) matrix with the BiMax algorithm, and we built the $\mathbf{T}_{B1}$ matrix with $p = 43$ *rows* and $m = 8$ *columns*. In order to choose the most

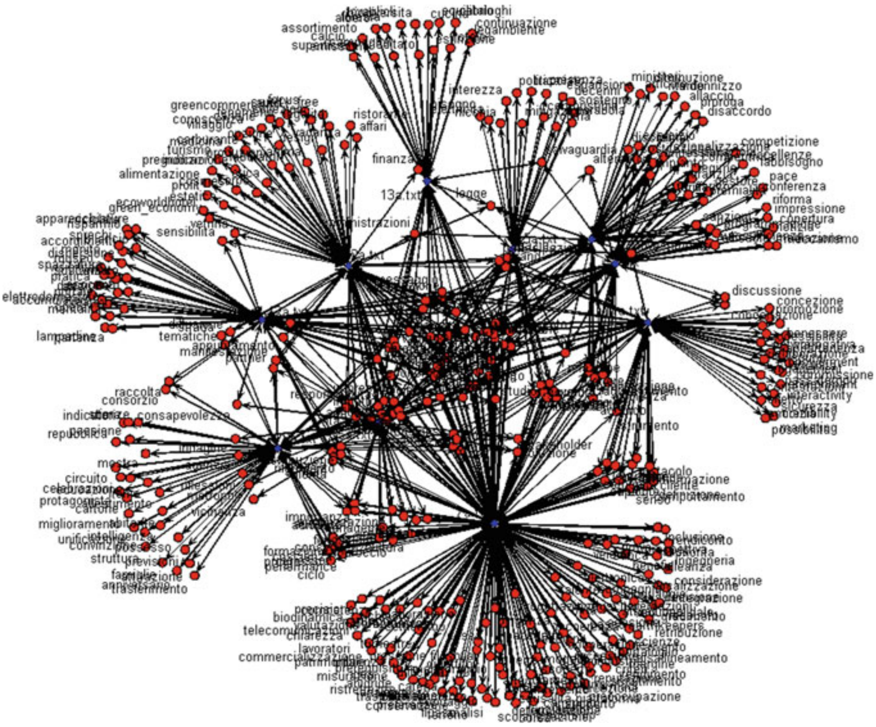


Fig. 3 Network N_{T_B} based on matrix $\mathbf{T}_B$ (*417*x*10*)

representative bipartition of terms and documents, 10(10 − 1) combinations of biclustering have been tested. The selected combination with $p = 2$ and $m = 2$, chosen on the highest value of Jaccard index, is composed by five biclusters. Overall, the biclustering algorithm has selected 43 terms in eight documents (matrix $\mathbf{T}_{B1}$).

Step Eight

We built the $\mathbf{W}_1$ (*43* × *43*) matrix. This matrix is a one-mode unweighted matrix, and it represents the relational system of the biclustering selected *textual form* (43).

Step Nine

We built the the heat map of the $\mathbf{W}_1$ matrix to visualize its sparsity (Fig. 4, *right side*).

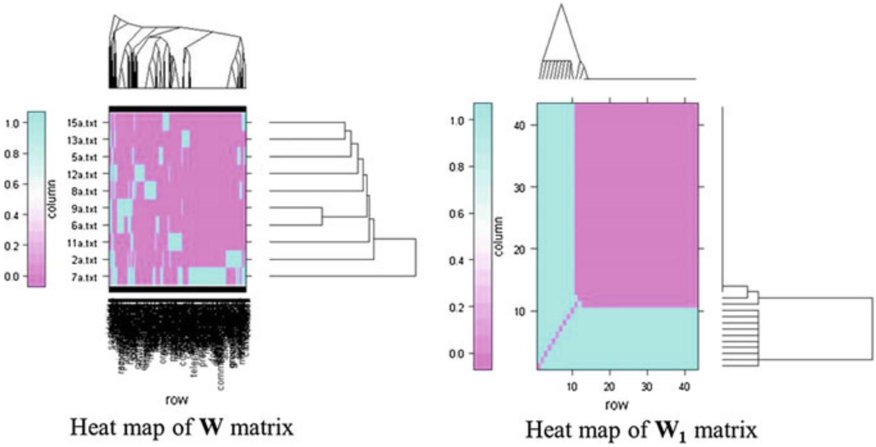
Step Ten

We analyzed the network N_{W_1} based on $\mathbf{W}_1$ matrix (Table 1, Figs. 5 and 6, *right side*).

Step Eleven

We compared the results for the network N_W and the network N_{W_1}.

First of all, we made a comparison between heat maps. Fig. 4 shows the $\mathbf{W}_1$ matrix (*43* × *43*) has much less sparseness than the $\mathbf{W}$ (*417* × *417*) matrix.



Heat map of **W** matrix

Heat map of $\mathbf{W}_1$ matrix

Fig. 4 Comparison between Heat maps

The analysis of the first network N_W leads to the formulation of an hypothesis of core–periphery structure. In the N_W, the core–periphery structure highlights that the core consists of 209 terms; this high value confirms the difficulty to identify a set of keywords able to spell out precisely the nature of the sustainability (Fig. 5). The search for boundary spanners has further confirmed the presence of a network structure characterized by a unique underlying meaning. More in detail, if the network terms involved different thematic areas, high values of betweenness centrality (C_B) would been found. Therefore, the C_B values of the first 20 terms indicate a unique meaning associated to the term *sustainability*, and C_B skewed distribution confirms this interpretation.

On the other hand, high values of closeness centrality C_C characterize not only the first 20 terms (more than 50% of the terms have very high values of closeness centrality), making it difficult to identify the terms which improve the meaning of *sustainability*. They do not clarify the link between terms, and the skewed distribution of C_C does not allow us to interpret the information (Fig. 6, *left side).*

Table 1 Descriptive statistics for networks: **W** and $\mathbf{W}_1$

Methods	W network	$\mathbf{W}_1$ network
Nodes	417	43
Links	62336	752
Avg. degree	149.48	18.27
Std. dev.	90.13	14.71
Density	0.30	0.41

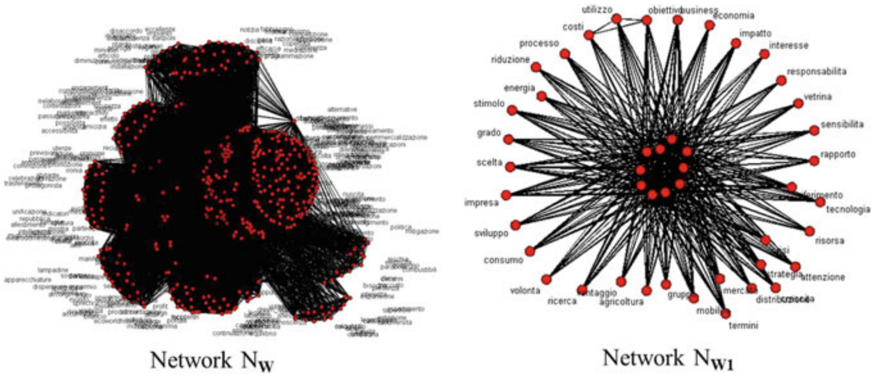


Fig. 5 Comparison between Networks

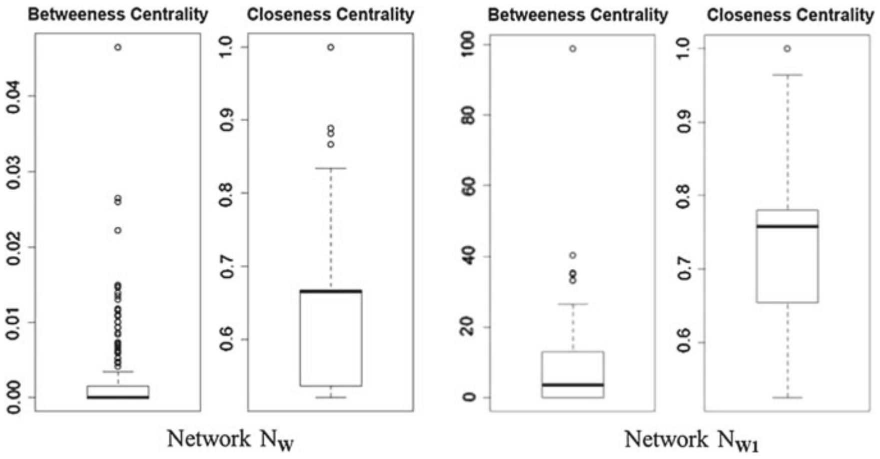


Fig. 6 C_B and C_C boxplots in **W** and $\mathbf{W}_1$

The descriptive statistics of the second network N_{W_1} show a global improvement of measures (Table 1, *right side*): The network is easily reading (with high variability of distribution degree), with a basket of keywords related to the sustainability meaning.

In details:

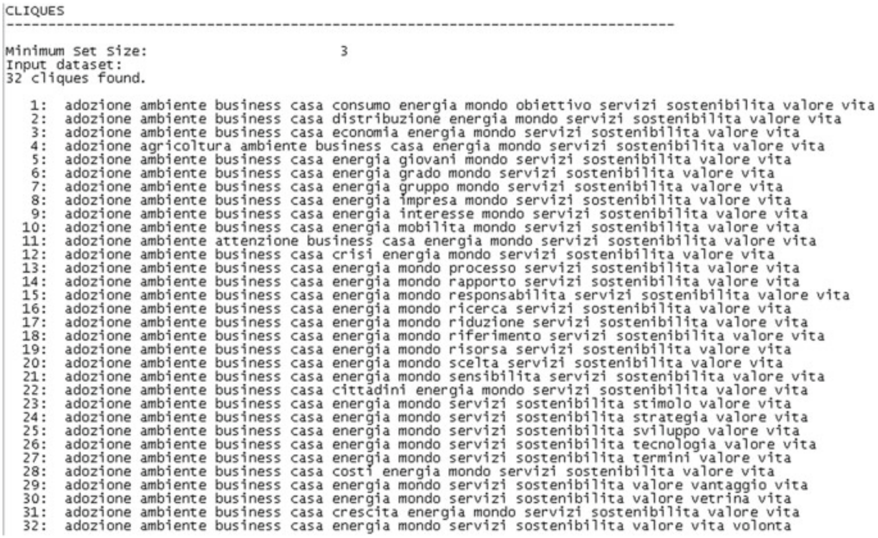
1. The density is equal to 0.41. This value shows a good cohesion of the textual network.
2. The N_{W_1} has a core–peripherical structure with several subgroups. The correlation coefficient equal to 0.972 shows that the core–periphery structure almost approaches the ideal one.
3. The terms in the *core* are as follows:
    - adoption,
    - environment,

- business,
- home,
- energy,
- world,
- services,
- sustainability,
- value, and
- life.

The network N_{W_1} presents the same data structure of N_W. The biclustering approach has preserved the initial core–periphery structure. The boxplots of the betweenness and closeness centrality in N_W and N_{W_1} show a drastic reduction of skewness (Fig. 7). The interpretation of N_{W_1} network is immediate, and the network is more readable.

The search for substructures through the analysis of cliques has highlighted 32 subgroups of concepts (Fig. 7) whose interpretation leads all the time to the environmental aspect. The composition of the subgroups presents for 94% of cases the following triads:

- adoption,
- environment,
- house.



```
CLIQUES
--------------------------------------------------------------------------------

Minimum Set Size:                       3
Input dataset:
32 cliques found.

   1:  adozione ambiente business casa consumo energia mondo obiettivo servizi sostenibilita valore vita
   2:  adozione ambiente business casa distribuzione energia mondo servizi sostenibilita valore vita
   3:  adozione ambiente business casa economia energia mondo servizi sostenibilita valore vita
   4:  adozione agricoltura ambiente business casa energia mondo servizi sostenibilita valore vita
   5:  adozione ambiente business casa energia giovani mondo servizi sostenibilita valore vita
   6:  adozione ambiente business casa energia grado mondo servizi sostenibilita valore vita
   7:  adozione ambiente business casa energia gruppo mondo servizi sostenibilita valore vita
   8:  adozione ambiente business casa energia impresa mondo servizi sostenibilita valore vita
   9:  adozione ambiente business casa energia interesse mondo servizi sostenibilita valore vita
  10:  adozione ambiente business casa energia mobilita mondo servizi sostenibilita valore vita
  11:  adozione ambiente attenzione business casa energia mondo servizi sostenibilita valore vita
  12:  adozione ambiente business casa crisi energia mondo servizi sostenibilita valore vita
  13:  adozione ambiente business casa energia mondo processo servizi sostenibilita valore vita
  14:  adozione ambiente business casa energia mondo rapporto servizi sostenibilita valore vita
  15:  adozione ambiente business casa energia mondo responsabilita servizi sostenibilita valore vita
  16:  adozione ambiente business casa energia mondo ricerca servizi sostenibilita valore vita
  17:  adozione ambiente business casa energia mondo riduzione servizi sostenibilita valore vita
  18:  adozione ambiente business casa energia mondo riferimento servizi sostenibilita valore vita
  19:  adozione ambiente business casa energia mondo risorsa servizi sostenibilita valore vita
  20:  adozione ambiente business casa energia mondo scelta servizi sostenibilita valore vita
  21:  adozione ambiente business casa energia mondo sensibilita servizi sostenibilita valore vita
  22:  adozione ambiente business casa cittadini energia mondo servizi sostenibilita valore vita
  23:  adozione ambiente business casa energia mondo servizi sostenibilita stimolo valore vita
  24:  adozione ambiente business casa energia mondo servizi sostenibilita strategia valore vita
  25:  adozione ambiente business casa energia mondo servizi sostenibilita sviluppo valore vita
  26:  adozione ambiente business casa energia mondo servizi sostenibilita tecnologia valore vita
  27:  adozione ambiente business casa energia mondo servizi sostenibilita termini valore vita
  28:  adozione ambiente business casa costi energia mondo servizi sostenibilita valore vita
  29:  adozione ambiente business casa energia mondo servizi sostenibilita valore vantaggio vita
  30:  adozione ambiente business casa energia mondo servizi sostenibilita valore vetrina vita
  31:  adozione ambiente business casa crescita energia mondo servizi sostenibilita valore vita
  32:  adozione ambiente business casa energia mondo servizi sostenibilita valore vita volonta
```

Fig. 7 Screenshot of UCINET for Clique

5 Conclusion

In this paper, we propose a strategy to derive an unweighted adjacency matrix from a lexical table (term–document matrix), after reducing its sparsity. We propose to apply a biclustering algorithm to extract the most significant information from the original term–document matrix and to build a new term–document matrix from which to derive an unweighted adjacency matrix for the network analysis.

The proposal strategy has been applied to the research of the common meaning of the term sustainability. The concept of sustainability does not have a specific and unambiguous definition; therefore, it lends itself to more than one interpretation, according to the contexts which it refer to (political, economic, or social). The analysis shows the network derived from the transformed lexical table has a stable structure and it is more simple: The core-periphery structure seems to be invariant, and the keywords seem to be aligned with the environmental context. The results are simple to read and useful for the interpretation of the true meaning of sustainability.

The main shortcomings are inherent use of BiMax biclustering algorithm that works only with binary matrix and the use of the quantile normalization for building the *one-mode unweighted* matrix $\mathbf{W}$: The first limit could be overcome by comparing other biclustering algorithms, and the second one could be overcome by testing other normalization methods. These limits represent the starting point for more in-depth analysis and its generalization.

References

Batagelj, V., & Mrvar, A. (2007). Hierarchical clustering with relational constraints of large data sets. In *Abstracts of the 6th Slovenian International Conference on Graph Theory, Bled*.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284*, 34–43.

Bolasco, S. (1999). Lánalisi multidimensionale dei dati, Carocci ed., Roma.

Borgatti, S. P., & Everett, M. G. (1999). Models of core/periphery structures. *Social Networks*, *21*(4), 375–395.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). UCINET for windows: Software for social network analysis.

Brundtland, G. H. (1989). *Sustainable development: An overview. Development*, *2*(3), 13–14.

Buchdahl, J. M., & Raper, D. (1998). Environmental ethics and sustainable development. *Sustainable Development*, *6*(2), 9298.

Butts, C. T. (2014). sna: Tools for social network analysis. R package version 2.3-2. http://CRAN.R-project.org/package=sna.

Carley, K. M. (1997). Network text analysis: The network position of concepts. In W. Roberts (Ed.), *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* (pp. 79–100). Mahwah, NJ: Lawrence Erlbaum Associates.

Carley, K. M., & Palmquist, M. (1992). Extracting, representing, and analyzing mental models. *Social Forces*, *70*, 601–636.

Corman, S., Kuhn, T., McPhee, R., & Dooley, K. (2002). Studying complex discursive systems: Centering resonance analysis of organizational communication. *Human Communication Research*, *28*(2), 157–206.

Danowski, J. A. (1993). Network analysis of message content. *Progress in Communication Sciences*, *12*, 198–221.

Davidson, K. M. (2011). Reporting systems for sustainability: What are they measuring? *Social Indicators Research*, *100*(2), 351–365.

Diesner, J., & Carley, K. M. (2005). Revealing social structure from texts: Meta-matrix text analysis as a novel method for network text analysis. In V. K. Narayanan & D. J. Armstrong (Eds.), *Causal mapping for research in information technology* (pp. 81–108). Harrisburg, PA: Idea Group Publishing.

Everett, M. G., & Borgatti, S. P. (2013). The dual-projection approach for two-mode networks. *Social Networks*, *35*(2), 204–210.

Feinerer, I., & Hornik, K. (2014). tm: Text mining package. R package version 0.5–10. http://CRAN.R-project.org/package=tm.

Franzosi, R. (1989). From words to numbers: A generalized and linguistics based coding procedure for collecting event data from newspapers. *Sociological Methodology*, *29*, 263–298.

Franzosi, R. (2010). *Quantitative narrative analysis*. Quantitative applications in the social sciences series: Sage.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239.

Giddings, B., Hopwood, B., & O'Brien, G. (2002). Environment, Economy and society: Fitting them together into sustainable development. *Sustainable Development*, *10*, 187–196.

Hartigan, J. (1974). BMDP3M: Block clustering. In W. Dixon (Ed.), *BMDP biomedical computer programs*. Berkeley: University of California Press.

Hunter, S. (2014). A novel method of network text analysis. *Open Journal of Modern Linguistics*, *4*(2), 350–366.

Imran, S., Alam, K., & Beaumont, N. (2011). Reinterpreting the definition of sustainable development for a more. *Sustainable Development*,. doi:10.1002/sd.537.

Leydesdorff, L., & Welbers, K. (2011). The semantic mapping of words and co-words in contexts. *Journal of Infometrics*, *5*, 469–475.

Madeira, S., & Oliveira, A. (2004). Biclustering algorithms for biological data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *1*, 24–45.

Middleton, N., O'Keefe, P., & Moyo, S. (1993). *Tears of the crocodile: From rio to reality in the developing world*. London: Pluto Press.

Murphy, P. E., & Price, G. G. (2005). Tourism and sustainable development. In W. F. Theobald (Ed.), *Global tourism* (Vol. 3, pp. 167–193). Burlington: Elsevier.

Novak, J., & Gowin, D. (1984). *Learning how to learn*. New York: Cambridge University Press.

Paranyushkin, D. (2011). *Identifying the pathways for meaning circulation using text network analysis*. Nodus Labs, Berlin: Technical Report.

Pearce, D. (1989). An economic perspective on sustainable development. *Journal of the Society for International Development*, *2*(3), 17–20.

Prelic, A., Preli, A., Bleuler, S., Zimmermann, P., Wille, A., Bhlmann, P., et al. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, *22*(9), 1122–1129.

Roberts, C. W. (1997). A generic semantic grammar for quantitative text analysis: Applications to East and West Berlin radio news content from 1979. *Sociological Methodology*, *27*, 89–129.

Sarvestani, A. A., & Shahvali, M. (2008). Environmental ethics: Toward an Islamic perspective. *American Eurasian Journal of Agricultural and Environmental Sciences*, *3*(4), 609–617.

Seghezzo, L. (2009). The five dimensions of sustainability. *Environmental Politics*, *18*(4), 539–556.

Sudhahar, S., Franzosi, R., & Cristianini, N. (2011). Automating quantitative narrative analysis of news data. *Journal of Machine Learning Research (JMLR)*, *17*, 63–71.

Sumudo, A. A. (2002). The right to a healthy life or the right to die polluted?: The emergence of a right to a healthy environment under international law. *Tulane Environmental Law Journal, 65*.

Tilbury, D., Stevenson, R. B., Fien, J., & Schreuder, D. (Eds.). (2002). *Education and sustainability: Responding to the global challenge*. IUCN, Cambridge, UK: Commission on Education and Communication.

Trigg, R., & Weiser, M. (1986). TEXTNET: A network-based approach to text handling. *ACM Transactions on Information Systems (TOIS)*, *4*, 1–23.

Vucetich, J. A., & Nelson, M. P. (2010). Sustainability: Virtuous or vulgar? *Bioscience*, *60*(7), 39544.

Wackernagel, M., & Rees, W. (1996). *Our ecological footprint*. Gabriola Island, Canada: New Society Publishers.

Wilkinson, L., & Friendly, M. (2009). The history of the cluster heatmap. *The American Statistician*, *63*(2), 179–190.

Zha, H., & Zhang, Z. (2000). Matrices with low-rank-plus-shift structure: Partial SVD and latent semantic indexing. *SIAM Journal on Matrix Analysis and Applications*, *21*(2), 522–536.

University of Bari's Website Evaluation

Laura Antonucci, Marina Basile, Corrado Crocetta, Viviana D'Addosio, Francesco D. d'Ovidio and Domenico Viola

Abstract Educational websites were studied from many different perspectives. In 2001, Zhang and von Dran developed a theoretical framework for evaluating website quality from a user satisfaction perspective, while Yoo and Jin in 2004 evaluated the design of university websites. In this paper, we assess the quality perceived by the users of the website of the University of Bari using factorial analysis and multiple correspondences analysis (MCA) visual maps. Latent variables resulting from this preliminary analysis were then used to evaluate the most important latent dimensions related to loyalty of the users. A segmentation analysis was performed to study how loyalty is influenced by variables and factors.

Keywords Customer satisfaction · University website · CATPCA · Factorial analysis · MCA · Classification tree

1 Framework and Survey's Description

The university websites are the most important information channel, in fact they provide general information, facilitate contacts between teachers and students, etc. Quality and usability of the websites are, therefore, very important to improve student satisfaction.

L. Antonucci · C. Crocetta
University of Foggia, Foggia, Italy

M. Basile
Language/Techno-Economic State High School "Marco Polo", Bari, Italy

V. D'Addosio
Professional State High School "Ettore Majorana", Bari, Italy

F.D. d'Ovidio (✉) · D. Viola
University of Bari Aldo Moro, Bari, Italy
e-mail: francescodomenico.dovidio@uniba.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_11

This work aims to evaluate the user satisfaction of the website http://www.uniba.it, using a ten-section CAWI questionnaire: *User profile*, *Graphics of the website*, *Website contents*, *Services*, *Error Handling*, *Website management*, *Interruptions management*, *Usability*, *Security/privacy* and, finally, *Overall Satisfaction*. The first nine sections contain several items, measured with a four- or five-level scale.

2 Explorative Analysis

Table 1 reports the average scores given by the 1,049 respondents to the main aspects considered, according to the frequency of access to the website. This frequency has an important role because it allows to distinguish occasional users from expert ones.

21.9% of respondents access the website only in few occasions, but 10.7% declare that they browse the website several times a day. 67.4% of respondents visit the website one to several times a week. In most cases students are quite satisfied, the average mark ranges from 3 to 4 in a five-point scale, and there are not great differences between occasional users and expert ones, but expert users are a little more satisfied than the others.

An exception concerns, obviously, the item "reporting of errors/malfunctions during browsing", because frequent users are presumably annoyed by errors/malfunctions more often than occasional users.

3 Identification of the Website Quality's Dimensions

The Bartlett's test of sphericity for the observed 46 items was very significant (p-value < 0.0000001), allowing the use of principal component analysis (PCA) to explore the dimensions of website's quality.

Because some observed variables are measured on few level categories and not normally distributed, the ALSOS CATPCA was applied instead of PCA .[1] By using a backward stepwise procedure, only factors with eigenvalues higher than 1.1 were selected, iteratively removing all items with communality lower than 0.51. As final result, we obtained a correlation matrix with 25 optimally scaled items, identifying six principal components that explain 70.2% of the overall variance.

[1] The CATPCA (categorical principal component analysis) algorithm is due to the Data Theory Scaling System Group of the Leiden University, NL (De Leeuw et al. 1976; Meulman et al. 2004). It belongs to the PRINCALS family, based on *Alternative Least Squares Optimal Scaling* procedures, allowing researcher to use categorical variables, while PCA requires at least interval-scaled variables and normal distribution of residuals. Incidentally, also classic PCA was performed in explorative way, providing almost the same results than CATPCA.

Table 1 Average rate of significant items, according to the user's frequency of access to the website of the University of Bari Aldo Moro; percentages of users access frequency

Statistically significant items* ($p < 0.001$)	Frequency of access				*All users*
	Never/at times	About once a week	Several times a week	Several times a day	
Utility level of the published information	3.36	3.55	3.71	3.64	*3.57*
Level of depth and detail of the content	3.07	3.15	3.22	3.32	*3.17*
Comprehensibility of the used lexicon	3.85	3.94	4.03	3.87	*3.94*
Reporting of errors/malfunctions during browsing	3.38	3.23	2.99	2.93	*3.16*
Duration of the service interruptions	3.05	3.06	3.10	3.09	*3.07*
Download time	3.65	3.78	3.91	3.85	*3.80*
Viewing the site on any browser	3.63	3.71	3.87	3.76	*3.75*
Appropriateness of the content discussion	3.42	3.64	3.49	3.62	*3.55*
Comprehensible and unambiguous terminology	3.39	3.57	3.63	3.67	*3.56*
User recognition	3.82	4.01	4.10	4.08	4.00
Overall assessment about the website	**3.42**	**3.51**	**3.48**	**3.71**	**3.50**
% by access frequency	21.9	37.2	30.2	10.7	100.0

*Statistics significances were obtained by using the test of maximum likelihood ratio ($\alpha = 0.05$)

The Kaiser-Meyer-Olkin value is very high (0.92), ensuring excellent fitting of the model to data.

Starting from the identified principal components, a factor analysis (Cattell 1952) was conducted by using non-orthogonal promax rotation, in order to obtain a simpler solution. The promax rotation allowed to identify the most characterizing variables for each latent dimension, preserving relationships between the factors (Manly 1986).

Table 2 shows the residual correlations not due to direct relationships among the observed items. Only the first four factors have high correlation coefficients showing a *structural relation* among factors.

In Table 3, the *communalities* column indicates the variability explained by the factorial system, or in other words, the importance of the observed item. The factor loadings express the intensity of the relationship between variables and factors.

Table 2 Correlation among factors in the promax solution*

Factors	F1	F2	F3	F3	F4	F5
F1	1	**0.480**	**0.628**	**0.434**	*0.270*	**0.397**
F2		1	**0.553**	**0.474**	0.089	*0.282*
F3			1	**0.496**	*0.187*	**0.343**
F4				1	0.084	*0.183*
F5					1	*0.104*
F6						1

*Statistical significance = Bold font: $p < 0.01$; Italic font: $p < 0.05$

Table 3 Factor loadings and communalities of the items of the promax rotated solution*

Items	Factors						*Communalities*
	F1	F2	F3	F4	F5	F6	
Clarity of the site map	0.949						*0.793*
Information's accessibility in a few clicks	0.918						*0.785*
Map accessibility	0.857						*0.696*
Categories classification while browsing	0.822						*0.705*
Understandable terminology	0.683						*0.555*
Useful information on the site	0.521		0.350				*0.604*
Services/activities simplification	0.482		0.366				*0.571*
Opening speed of the pages		0.910					*0.839*
Website load speed		0.908					*0.815*
Download speed		0.879					*0.776*
Scrolling speed		0.836					*0.758*
Viewing the site on every browser		0.827					*0.705*
Comprehensibility of the used lexicon			0.868				*0.699*
Utility of the published information			0.849				*0.703*
Clarity of the contents			0.809				*0.730*
Level of depth and detail of the content			0.795				*0.713*
Adequacy of the contrast between font and background colour				0.855			*0.776*
Font size				0.808			*0.733*

(continued)

Table 3 (continued)

Items	Factors						*Communalities*
	F1	F2	F3	F4	F5	F6	
Visibility of the website features				0.712			*0.556*
Language selection					0.890		*0.760*
Responsiveness/alerts of technical inefficiency in the contact form					0.789		*0.633*
Accuracy/correctness of the translation					0.648		*0.606*
Error messages/corrective action						0.891	*0.811*
Alerts of errors or malfunctions						0.785	*0.640*
Error/data recovery						0.659	*0.593*

*Factor loadings lower than 0.33 have been omitted in this table

By evaluating such relationships, the factors can be then interpreted as follows:

- Factor 1: Accessibility and usability;
- Factor 2: Access speed;
- Factor 3: Information and content;
- Factor 4: Graphics and readability;
- Factor 5: Interactions;
- Factor 6: Error handling.

4 Proximity Map of the Observed Items

In order to confirm factorial similarities and to identify the main relationship, a visual map was used. Figure 1 shows the first two dimensions resulting from the multiple correspondence analysis obtained using the ALSOS algorithm: HOMALS (De Leeuw and Van Rijckevorsel 1980).

The position of the 25 centres of gravity of the observed variables highlights the relationships among the factors to which these variables are related (de Leeuw 1984; Gifi 1990). The points related to each factor are inserted in a shape with the corresponding number of the factor.

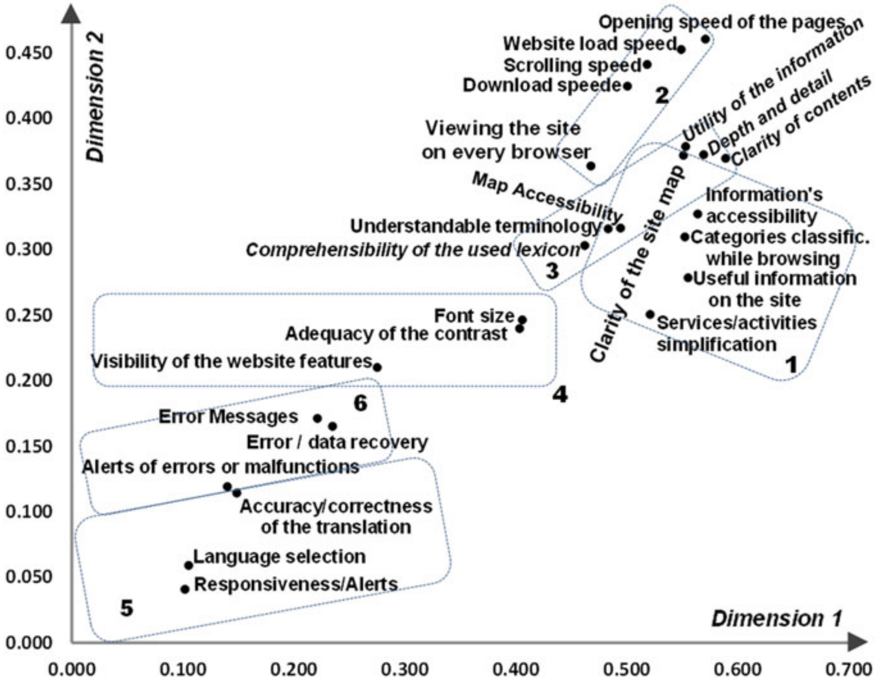


Fig. 1 Multiple correspondences map of observed items (first two dimensions)

The first two dimensions of MCA explain more than 70% of the total inertia. Figure 1 shows that the results of the factorial analysis are quite congruent with the two dimensions of the MCA.

The centres of gravity are concentrated along the main diagonal, ranking variables, and factors according to their importance with respect to the unidimensional concept of quality. The lower end of the diagonal (the less important items) is identified by the variables corresponding to factor "interactions", while the factor "access speed" identifies its upper end, i.e. the most important variables.

5 Quality Dimensions and Loyalty Elements

Loyalty can be predicted through classification methods. After many attempts, we choose to try a classification tree using the binary variable "access frequency" as response, where *high frequency* grouped the answers "several times a week" and "several times a day", while *low frequency* was associated with the other answers.

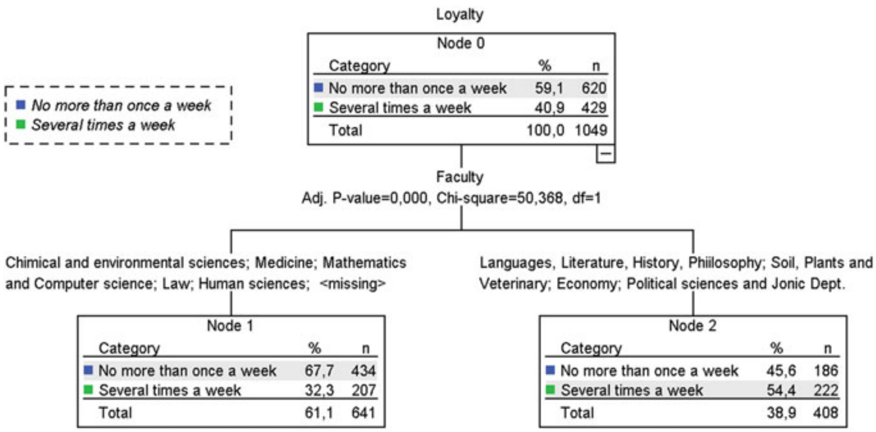

Loyalty
Node 0
Category % n
🟦 No more than once a week 59,1 620
🟩 Several times a week 40,9 429
Total 100,0 1049
🟦 No more than once a week
🟩 Several times a week
Faculty
Adj. P-value=0,000, Chi-square=50,368, df=1
Chimical and environmental sciences; Medicine; Mathematics and Computer science; Law; Human sciences; <missing>
Languages, Literature, History, Phiilosophy; Soil, Plants and Veterinary; Economy; Political sciences and Jonic Dept.
Node 1
Category % n
🟦 No more than once a week 67,7 434
🟩 Several times a week 32,3 207
Total 61,1 641
Node 2
Category % n
🟦 No more than once a week 45,6 186
🟩 Several times a week 54,4 222
Total 38,9 408

Fig. 2 Classification tree to predict the frequent access to the UNIBA website

All the interviewees characteristics (gender, residence, faculty, etc.) were selected as predictive variables, as well as the six quality factors identified above.[2]

The best known classification methods, CRT (Breiman et al. 1984) and CHAID (Kass 1980), were used, fixing 30 cases as minimum frequency of child nodes, expanded on maximum five levels of classification, and assessed by using cross-validation with 25 subsamples.

The chosen model, performed by using CHAID, can correctly predict the 62.5% of cases according the Faculty/Department (Fig. 2). The classification tree points out that students attending humanistic courses use the website more often than their colleagues of scientific courses.

The quality factors (precisely, "access speed", "information and content", and "interactions") appear at the second and third level of the classification tree, but without any effect on the predictive power of the model and thus they were removed by manual pruning.

The outcomes for the two cases "not more than once a week" and "several times a week" are quite different (see Table 4), because the latter response seems to be more difficult to identify.

The results here obtained are very good and robust, given that crossvalidation provides exactly the same risk values than the main classification (Table 5).

[2] The user's evaluation of the website could influence the frequency of access, because satisfied users tend (*cæteris paribus*) to browse the site more often than unsatisfied ones.

Table 4 Confusion matrix (classification table)

Observed website access frequency	Predicted website access frequency		
	Not more than once a week	Several times a week	Correct classification (%)
Not more than once a week	434	186	***70.0***
Several times a week	207	222	***51.7***
Total (%)	*61.1*	*38.9*	***62.5***

Table 5 Risk table

Method	Risk estimate	Std. error
Resubstitution	0.375	0.015
Crossvalidation	0.375	0.015

6 Concluding Considerations

This study showed a hierarchy of the variables, connected to the six dimensions of quality. Among them, the technical dimensions ("accessibility and usability" and "access speed") seem to be the most important, while the main mission of a website (providing *information and content*) has only the third position.

These findings were used, in addiction to the interviewees' characteristics, to analyse variables with respect to the loyalty proxy "access frequency to the website", by using segmentation analysis. Only a strong Faculty/Department effect was found, and this appears logical because, as it is known, the services are usually provided by these institutions following rules fixed at central level.

The main conclusion of this study is that the website quality has a weak influence on the "users loyalty", despite the current opinion "the higher the quality, the higher the loyalty".

Certainly, the analysis of the websites quality can not be limited to the few aspects described in the previous pages. This study should be considered just a first approach to the problem. Further analyses can start by the structural relationships here found among the quality dimensions, in order to find a causal model able to better explain the user behaviour.

Acknowledgements The authors jointly designed and realized the work here described. However, L. Antonucci wrote the Sect. 5, M. Basile wrote the Sect. 3, C. Crocetta wrote the Sect. 6, V. D'Addosio wrote the Sect. 2, F. D. d'Ovidio wrote the Sect. 4, and D. Viola wrote the Sect. 1. The Authors thank the University of Bari Aldo Moro (Italy) that provided the data. The usual disclaimer applies.

References

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. New York-London: Chapman & Hall.

Cattell, R. B. (1952). *Factor analysis*. New York: Harper.

de Leeuw, J. (1984). *Canonical Analysis of categorical data* (2nd ed.). Leiden (NL): DSWO Press.

De Leeuw, J., & Van Rijckevorsel, J. (1980). Homals and princals—Some generalizations of components analysis. In: E. Diday, Y. Escoufier, L. Lebart, J. P. Pages, Y. Schektman, R. Tomassone (Eds.), *Data analysis and informatics* (pp. 231–241). Amsterdam, NL.

de Leeuw, J., Young, F. W., & Takane, Y. (1976). Additive structure in qualitative data: An alternative least squares method with optimal scaling features. *Psychometrika, 41,* 471–504.

Gifi, A. (1990). *Nonlinear multivariate analysi*s. Wiley.

Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics, 29*(2), 119–127.

Manly, B. F. J. (1986). *Multivariate statistical methods: A primer* (p. 77). London: Chapman & Hall.

Meulman, J. J., Van der Kooij, A. J., & Heiser, W. J. (2004). Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 49–70). London: Sage.

Yoo, S., & Jin, J. (2004). Evaluation of the home page of the top 100 university web sites. *Academy of Information and Management Sciences, 8*(2), 57–69.

Zhang, P. & von Dran, G. M. (2001). Expectations and ranking of website quality features: Results of two studies on user perceptions. In *Proceedings of the 34th Hawaii International Conference on System Sciences (HICSS34)*, January.

Advantages of Administrative Data: Three Analyses of Students' Careers in Higher Education

Andrea Amico, Giampiero D'Alessandro and Alessandra Decataldo

Abstract This paper focuses on the potentiality of transformation of cross-sectional administrative information into longitudinal sociological data. The case study focuses on the Sapienza University of Rome administrative archives and the strategies of extrapolation and management of its registered students' information in order to obtain a longitudinal data set. Furthermore, this paper shows the analysis of this longitudinal data set in order to highlight the utility of this kind of data structure and with an aim to: (1) evaluate Italian Higher Education reform policies and their implementation outcomes with reference to their main purposes through a quasi-experimental research; (2) prove that the Sequence Analysis (SA) tools are suitable for studying the complexity of Italian higher education students' careers; (3) analyse the factors (also contextual) that have had an impact on the success/failure throughout the students' academic careers (through Event History analysis, EH).

Keywords Administrative data management · Longitudinal studies · Quasi-experimental research · Sequence Analysis · Event History analysis

A. Amico · G. D'Alessandro
Department of Comunicazione e Ricerca Sociale, Sapienza University of Rome,
via Salaria 113, 00198 Rome, Italy
e-mail: andrea.amico@uniroma1.it

G. D'Alessandro
e-mail: giampiero.dalessandro@uniroma1.it

A. Decataldo (✉)
Department of Sociologia e Ricerca Sociale, University of Milano Bicocca,
via Bicocca degli Arcimboldi 8, 20126 Milan, Italy
e-mail: alessandra.decataldo@unimib.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_12

1 Introduction

In the fields of information technology and computational sciences, the debate (Buhl et al. 2013) on the use of big databases has been overcome by the developments in the DBMS (DataBase Management System). In social sciences, the issue is set in a completely different way (Manovich 2011; González-Bailón 2013). Even when the researchers have access to a large amount of information, it cannot be directly used for the purpose of a research. The vast majority of the data gathered and recorded each day are not designed for social research aims, but for the administrative, business or communicative purposes. It is the task of a researcher to choose the information of the greatest relevance and the highest quality, and to organize the database in an optimal way. Only after this search phase, it is possible to deal with information extrapolation and the phases of organization, quality control, cleaning, encoding and integration of the data (we refer to all these phases as management). Underestimating the importance of the search, extrapolation and management phases might jeopardize the results of the whole research.

This paper focuses on the potentiality of transformation of cross-sectional administrative information into longitudinal sociological data. Throughout this procedure, it is possible to investigate the same data set in various ways in order to reach different goals. In this paper, we discuss three different analyses that were conducted on the identical data set with (partially) different aims.

2 Research Aims

Italian university system has been characterized by the radical reforms introduced by the “Bologna process”. Previously, this system was characterized by three aspects (Benvenuto et al. 2012): (1) low number of graduates; (2) excessively long university careers; (3) very high number of drop outs (source: CNVSU 2006). The AY 2001–2002 represents the transition from an university education based on a single level degree (4, 5 or 6 years) to a dual level degree: bachelor—BA (3 years)/master—MA (2 years) or other types of master (5 or 6 years, such as Architecture or Medicine). The principal aim of this reform was to control the phenomena of dropping out, perpetual students and low number of graduates.

In the course of the last decades in Italy, the main agencies that have dealt with the university system evaluation used cross-sectional and repeated over-time analyses. These analyses are used in order to: (1) evaluate the reform achievements; (2) establish the Italian university standards; (3) allocate the public funding among university institutions. Therefore, the lack of longitudinal data set related to the students’ careers prevents an adequate evaluation.

We identified Sapienza University of Rome as a representative case because it is the largest university in Italy (for more details, see: https://en.wikipedia.org/wiki/List_of_universities_in_Italy).

Firstly, our work focuses on the Sapienza administrative archives and the strategies of extrapolation and management of its registered students' information in order to obtain a longitudinal data set. Secondly, this paper shows the analysis of this longitudinal data set in order to highlight the utility of this kind of data structure and with an aim to: (1) evaluate Italian Higher Education reform policies and their implementation outcomes with reference to their main purposes through a quasi-experimental research; (2) prove that the SA tools are suitable for studying the complexity of Italian higher education students' careers; (3) analyse the factors (also contextual) that have had an impact on the success/failure throughout the students' academic careers (through EH).

3 Data

Information concerns the cohorts of the enrolled students before (from AY 1991–1992 to AY 2000–2001) and after the reform (from AY 2001–2002 to AY 2012–2013). The pre-2001 cohorts were monitored until March 2008 and the post-2001 cohorts until March 2014. Therefore, the first cohort was followed for a period of 16 years, the second one for 15 years, etc.

This information is gathered for administrative purposes, such as monitoring the fee payments and the examination registrations. However, they are collected in different archives. They relate to four main areas: socio-economic profile, high school education, higher education career and university performance indicators. This kind of (de)structuration is not suitable for the purpose of temporal dynamics studying. Furthermore, the information is registered in a cross-sectional way (each information refers to a single moment), and this represents a further critical issue for temporal dynamics studying. Indeed, the phenomena characterized by continuous process of change can be studied in a more appropriate way through the use of longitudinal data (Blossfeld et al. 1989).

Starting from these archives, a unique data set has been re-constructed in a longitudinal way, using key codes (the identification numbers of the study course and of the students). Through the recording of the events related to each student's career (payment of every semester fees from the first enrolment to graduation or dropping out, passing of the examinations, etc.), a longitudinal data vector for each student was created. By positioning the events on a chronologically ordered string, it is possible to control the dynamics of each academic career and to highlight specific phenomena (dropping out, stopping out, mobility, degree attainment, etc.). These data allow us to follow students' careers from the first enrolment until graduation (or dropping out).[1]

[1] The data query was problematic and it required over 1,000 lines of Structured Query Language (SQL) code.

In the final data set (544,025 students) the data referring to the socio-demographic characteristics and high school training are cross-sectional (registered at students' first enrolment), whereas those referring to university career and performance indicators are longitudinal (repeated for each semester).

4 Methodology

4.1 The Quasi-experimental Design

The first study (Benvenuto et al. 2012) aims to evaluate the Italian Higher Education reform policies and their implementation outcomes with reference to their main purposes through a quasi-experimental logic of investigation (Shadish et al. 2002). Through the adoption of this logic, we are able to attribute to the university reform (which is considered as the experimental variable), the differences between the outcomes of the post-2001 cohorts (the post-test population) and the ones of the pre-2001 cohorts (the pre-test population).[2] The research involves the cohorts of students enrolled from AY 1991–1992 to AY 2007–2008.

All the students (413,336) are monitored for a period that doubles the length of their study course (i.e. the BA students are monitored for 6 years). This allows us to compare: (1) careers of the students enrolled in course with different length; (2) students' careers belonging to different educational systems (pre- and post-2001). Thereby, the aim is to analyse the students' careers by identifying the characteristics of their academic trajectories (such as years of stopping out) related to success/failure. This requires taking into consideration several variables referring to input, throughput and output career aspects. The final typology (a combination of all these variables) is a very complex descriptive system with 966 different types of careers. Through this typology, we can compare the outcomes of pre-2001 cohorts with those of post-2001 ones in order to evaluate the impact of the reform.

4.2 Sequence Analysis for the Exploration of Students' Careers

The second study uses the SA; it is an efficient method that involves the examination of ordered social process. We use SA in order to (1) describe in detail the phenomena of late graduation and late performance (retention); (2) identify different

[2] With reference to the classical experimental approach by Campbell and Stanley (1966), the adopted research design represents a combination of the 7th design (Time-Series experiment) and the 12th design (Separate-Sample, pre-test and post-test).

kinds of students dropping out from the university (attrition); (3) evaluate other particular phenomena that could delay the careers (such as the mobility within and between faculties).

The idea is to list each student's academic trajectories at Sapienza as an ordered list of administrative states, which are defined on a time axis. The time interval being used is the semester. Due to the relevance of the AY 2001–2002, the analysis focuses only on this cohort (23,854 students). Each student is monitored up to 25 semesters.

For the scope of this analysis, we consider ten different administrative "states" that define the alphabet for the SA. They are: (1–4) four types of enrolment (due to the course length: 2, 3, 5 or 6 years); (5) continuation (after the previous semester); (6) graduation; (7) formal dropping out (with a request of study interruption, probably due to a transition to another university); (8) informal dropping out (due to no six-month fee payment); (9) change of faculty (between two faculties); (10) change of course (within the same faculty).

A distance matrix between each pair of careers was calculated by Optimal Matching (OM) algorithm (Studer and Ritschard 2014; Blanchard et al. 2014). The OM produced a huge matrix, used for Wald cluster analysis. These six groups are used as dependent variables in a multinomial logistic regression model (ML). The purpose is to achieve well-defined clusters and to characterize them with reference to the information regarding the students, the first year performance indicator and the university context.

4.3 Event History Models: Success or Failure Factors

The third study focuses on EH models. Through the implementation of EH, it is possible to study simultaneously the changes occurred over a particular period of time in various contexts and in different actors (Blossfeld et al. 1989). The EH allows precisely to relate the duration and the performance of the processes with the independent variables or covariates that are theoretically relevant. It is, therefore, possible to make inferences about the influence of covariates on the length and on the occurrence (or non-occurrence) of an event.

In this work, the events that were analysed are mainly two: graduation and dropping out. The probability that one of these two events occurs in a certain moment of time is the core of this analysis (Box-Steffensmeier and Bradford 2004). Firstly, this study aims to identify the most suitable model of EH that could respond to the specific cognitive needs. Only after the right model (the one with the shape of the hazard function distribution that better fits to the data) was identified, the personal and contextual covariates were introduced. In the final log-logistic parametric survival model, the administrative data of the careers were integrated with context indicators related to the organizational structure of the university and with some time-related covariates referring to the socio-economic situation. This study supports the idea that some external conditions (such as low receptivity of the labour market and low rates of overall economic growth) could affect the decision of a student to use an university as a *parking place*, even for a longer period, before attempting to enter in the labour market (Brunello and Winter-Ebmer 2003; Häkkinen and Uusitalo 2003). In order to allow the comparison among different university careers, this study is restricted only to the students enrolled from AY 2001–2002 to AY 2010–2011 in BA courses (200,554 students).

5 Results

5.1 The Quasi-experimental Study

The principal outcomes show that, on the one hand, the transition from the old to the BA/MA university system has a positive impact with regard to the regular graduate rate. It passed from 2% to nearly 16% average in the post-2001 cohorts. On the other hand, the reform does not have the same impact on the late graduate rate. It increased from 28.6% of the last cohort of the old system to 34.6% of the first cohort of BA/MA system. However, this percentage decreased in the last cohorts that were monitored.

The reform did not affect the amount of students that were still enrolled at the double of their course length, which was decreasing even in the last pre-reform cohorts.

With regard to the dropping out rate, although the reform caused a 15% leap, it remained very high. The retention of the system has improved, but still the performance was not adequate, considering that more than 2/5 of students were dropping out from the university in the post-2001 cohorts.

These four administrative statuses include very different trajectories, such as careers with stopping out episodes or with mobility events. The typology describes carefully these trajectories, but it does not permit to understand the connections among the events of each trajectory and to relate them with students' outcomes. For this purpose, the SA is useful.

5.2 The SA

The SA approach has proven to be valuable for pointing out the relevance of phenomena in temporal order (Blanchard et al. 2014). This analysis is an efficient way to study the phenomena of late graduation, late performance, dropping out and other particular aspects (such as within and between faculties mobility) that could have impact on the length of the student's career. The main hypothesis is that the *timing* (i.e. the observation point when an event occurs) is the most relevant aspect which determinates the career outcome.

Through cluster analysis, performed on the OM distance matrix, six well-defined groups of students were identified. The first group consists of 2,926 (12.3%) students. In this group, the majority (35.5%) is composed of students that are still enrolled at the end of the observation period (25th semester); other part of the group are the drop outs (informal, 32.3%, and formal 8.3%) or graduates after 10 years since the enrolment (22.2%). The second (2,858 students, 12% of total number) and the fourth groups (4,984 students, 20.9%) differ for the dropping out timing: the students belonging to the second group keep postponing the dropping out decision until the ninth semester, whereas for the fourth group the dropping out occurs in the first half of the observation period. The second type of dropping out (the formal one) characterizes the sixth group (3,574 students, 15%). In other words, the majority of these students drops out at the beginning of their career. The other two groups (the third and the fifth) are composed of the students that end their career with graduation. Although the students of the third group (4,470, 18.7%) graduate at the regular course length, the timing of graduation in the fifth group (5,042 students, 21.1%) is slower.

These six groups are used as dependent variables in a ML. It includes two types of independent variables: micro level—social characteristics of students (sex, age at enrolment, residence and family income), high school information (high school type and high school final mark), and first year performance indicators (credits and average mark); meso level—university context information (length of the first enrolment study course and the limited enrolment course).

The main result shows the importance of the *warming-up* period in determining the development of each career. The first year performance indicators are the best predictors of success. Furthermore, the university context information, especially the enrolment course length, is crucial: students enrolled in a 5- or 6-year course have more chances to graduate than BA students.[3]

In this analysis, we used only the information strictly related to the students (micro level) and the university context (meso level). Due to the importance of the meso level, it seems interesting to include also the external context (macro level). For this purpose, the EH is useful.

5.3 The EH

In order to understand whether and how the meso and macro context affects the students' careers, the administrative (micro) data of students were integrated with

[3] In order to study the relevance of the connections among the events of each trajectory (such as within and between mobility), the OM analysis was performed on the careers with one or more mobility events (4,804 students, 20.1% of the enrolled in 2001–02 cohort). The outcome confirms the relevance of the *warming-up* period (in terms of study course reorientation and first year performance) in determining the success/failure of each career.

meso (faculty area, stationary condition[4]) and macro contextual data (unemployment rates at provincial and age related level, variation of GDP). Owing to the longitudinal construction, it is possible to treat the macro contextual data as time-varying covariates. This allows to understand whether and how the variation in the economic context has the influence on the careers.

The students with an age between 20 and 24 years at the enrolment have the lower risk of reaching the event (graduation) in a short period (49% lower than the younger ones). Students with higher high school final mark and with scientific high school type have also more probability to obtain a degree within a lower amount of time. Students with family residence outside Lazio (the Sapienza region) have 29% higher probability to have a short and successful university career. Low and middle family income affect duration of the careers in the same direction (by accelerating them).

Meso (structural) covariates are mainly referred to the faculty of registration. The remarkable differences between faculty areas are attributable to the strict enrolment selection of medic courses and to the different presence of the students with top high school final marks (Benvenuto et al. 2012). The outcome shows clearly that a stationary career is more successful and faster.

The contextual external covariates confirm the initial hypothesis. There is a lower risk of obtaining a degree with high unemployment rates during the first two years of the career, and there is a higher risk when the GDP is rising. According to the initial hypothesis, this model shows that some external conditions (low receptivity of the labour market, lower overall economic growth rates and higher disposable income) affect student's decision to use the university as a *parking place* before attempting to approach the labour market.

6 Summary and Conclusions

This paper shows the value of transforming cross-sectional administrative information into longitudinal sociological data vectors. The longitudinal data set allows us to investigate students' careers dynamics throughout appropriate analysis in order to reach various aims. The evaluation of the Italian university reform (the first aim) is reached through a quasi-experimental design: the outcomes show that the reform was a partial success. The students' careers complexity (the second aim) was examined in detail with the use of SA: the main results show the importance of the *warming-up* period. The impact of contextual meso and macro factors (the third aim) was studied through MLs and EH: in particular, the last one shows the relevance of some external conditions that could affect student's decision to use the university as a *parking place*.

[4] The stationary condition is referred to the comparison between the field of studies of the last registration course and the one at which the students enrolled at the beginning.

In conclusion, without precise theoretically oriented and methodologically adequate phases of extraction and management, the administrative information is not so expendable outside the bureaucratic goals. On the contrary, if adequately managed, these data may have benefits in evaluation studies, especially in benchmark research among different study courses, faculties, an entire university or national university system. Furthermore, the results can be immediately used to implement the improving policies. For example, once the *warming-up* period has been identified as the most important in a student's career, Sapienza (or other University) could implement a more effective tutoring activities during the first year (or semester) of a course.

References

Benvenuto, G., Decataldo, A., & Fasanella, A. (Eds.). (2012). *C'era una volta l'università? Analisi longitudinale delle carriere degli studenti prima e dopo la "grande riforma"*. Acireale–Roma: Bonanno.

Blanchard, P., Bühlmann, F., & Gauthier, J. A. (Eds.) (2014). *Advances in sequence analysis: Theory, method, applications. Series: Life course research and social policy,* (Vol. 2). New York, NY: Springer.

Blossfeld, H. P., Hamerle, A., & Mayer, K. U. (1989). *Event history analysis*. Hillsdale, NJ: Erlbaum.

Box-Steffensmeier, J. M., & Bradford, S. J. (2004). *Event history modeling: A guide for social scientists*. Cambridge, UK: Cambridge University Press.

Brunello, G., & Winter-Ebmer, R. (2003). Why do students expect to stay longer in college? Evidence from Europe. *Economic Letters, 80*(2), 247–253. doi:10.1016/S0165-1765(03)00086-7.

Buhl, H. U., Röglinger, M., Moser, F., & Heidemann, J. (2013). Big data a fashionable topic with (out) sustainable relevance for research and practice? *Business and Information Systems Engineering, 5*(2), 65–69. doi:10.1007/s12599-013-0249-5.

Campbell, D. T., & Stanley, J. (1966). *Experimental and quasi-experimental design for research*. Boston, MA: Houghton Mifflin Company.

CNVSU—National Committee for the Evaluation of the university system. (2006). Settimo Rapporto sullo Stato del Sistema Universitario. Retrieved August 24, 2016, from http://www.cnvsu.it/_library/downloadfile.asp?id=11341

González-Bailón, S. (2013). Social science in the era of big data. *Policy and Internet, 5*(2), 147–160. doi:10.1002/1944-2866.POI328.

Häkkinen, I., & Uusitalo, R. (2003). *The effect of a student aid reform on graduation: A duration analysis* (2003:8). Working Paper. Department of Economics, Uppsala: Uppsala University.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. In M. K. Gold & L. F. Klein (Eds.), *Debates in the digital humanities* (pp. 460–475). Minneapolis, MN: University of Minnesota press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

Studer, M., & Ritschard, G. (2014). A comparative review of sequence dissimilarity measures. *LIVES Working Papers, 33*, 1–47. doi:10.12682/lives.2296-1658.2014.33.

Growth Curve Models to Detect Walking Impairment: The Case of InCHIANTI Study

Catia Monicolini and Carla Rampichini

Abstract This work is motivated by a study on aging based on a representative population living in the Chianti geographic area (Tuscany, Italy). Multiple factors may influence the ability to walk, and no standard criteria are currently available to establish whether these factors are functioning within the "normal" range. Our work exploits data collected during the performance of an indoor mobility test to discriminate individuals at high risk of mobility disability and/or falls. A large number of outcomes were collected during the test. We explore the application of growth curve models to detect walking impairment. The model is extended to include nonignorable missing data. Our findings suggest that subjects aggregate into distinct behavioral profiles, characterized by age and other demographic and anthropometric factors.

1 Introduction

Our sample is from the InCHIANTI Study,[1] a representative population-based study of older persons living in the Chianti geographic area (Tuscany, Italy).

The main purpose of the study is to translate epidemiological research into geriatric clinical tools that make possible more precise diagnosis and more effective treatment in older persons with mobility problems.

The baseline study was conducted in 1998 on 1453 subjects from Greve in Chianti and Bagno a Ripoli.

We consider a subsample from the last follow-up (2013–2014): 316 subjects aged 35–93 years that performed the 400-m walk test. This test is well known in literature. It is predictive of cardiovascular, mental, cognitive, musculoskeletal, and neurological problems, and it recognizes psychophysical difficulties that are not shown by

[1] http://inchiantistudy.net/wp/.

C. Monicolini · C. Rampichini (✉)
Department of Statistics, Computer Science, and Applications 'G. Parenti',
Università degli Studi di Firenze, Florence, Italy
e-mail: rampichini@disia.unifi.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_13

other tests (Tian et al. 2015a, b; Vestergaard et al. 2009; Sayers et al. 2006; Newman et al. 2006; Chang et al. 2004).

The InChianti test is developed in collaboration with the FARSEEING EU project[2]: subjects perform the test with a smartphone located next to the last vertebra of their body, and the smartphone is equipped with an accelerometer collecting movement data, transformed into gait parameters. The subjects were instructed to walk at their maximum speed, wearing everyday shoes. Since the total run of 400 m was split into 44 straight of 9.09 m, the parameters of interest are registered many times for each subject; thus, the data have a repeated measures structure. Our analysis focuses on the most relevant parameters to detect walking problems, exploiting growth curve models to properly take into account repeated measures and subject heterogeneity (Hedeker and Gibbons 2006).

2 Data

The data set contains repeated test measures alongside with many individual characteristics collected before the walk test, by a questionnaire, a medical examination, and a session of mobility and mental performance. For every straight a subject complete, the walk time and the data from the smartphone were collected. If a subject does not complete the walking test, his parameters values are missing just after the last straight completed. Since the gait of a subject has particular characteristics at the beginning of the test (just started walking) and at the end (the subject know the test was almost over), the first two and last two straights are not considered in the analysis. There are 26 parameters collected during the walking test; the most relevant are the following:

- **Straight walk time**: a simple measure of walking problems;
- **Variation Coefficient of Cadence**: calculated as 100 × SD(gait time), where high values of this measure point out walking problems and low coordination;
- **Harmonic Ratio**: describing movement's smoothness, calculated with respect to 3 axes (antero-posterior, AP; medio-lateral, ML; vertical, V); this measure discriminate young from elder subjects;
- **Normalized Jerk Score**: recently proposed by Palmerini et al. (2013) to describe movement's smoothness, calculated with respect to 3 axes (AP, ML, V);
- **Symmetry Index**: describing the symmetry within right and left step. 0 stays for perfect symmetry, while higher values detect asymmetry; it can be calculated with respect to 3 axes (AP, ML, V).

Table 1 reports simple statistics of the pre-test subject characteristics and of the test measures used in the subsequent analysis.

[2] http://farseeingresearch.eu/.

Table 1 Variables description

Variable	Average	Std dev	Min	Max
Pre-test covariates				
Age	69.9335	15.7940	35	93
Physical activity	1.9399	0.5896	1	3
Height (cm)	162.0571	9.7800	135	188
Subject average of repeated test measures				
APJ score	0.9003	0.2931	0.1695	2.9811
AP harmonic ratio	2.1389	0.4286	0.6824	3.4776
AP simmetry index	0.1276	0.2271	0.0168	2.0819
Var.Coef. of Cadence	5.6864	3.8319	2.1469	39.1506
SD(simAP)	0.0844	0.1449	0.0121	1.2523
Dropout indicator				
D (1 if subject stops before completion)	0.0949	0.2931	0	1

3 Analysis of Antero-Posterior Jerk Scores

The analysis focuses on the Normalized Jerk Scores, since, for our knowledge, this is the first time they were collected as longitudinal measures. In particular, we consider the Antero-Posterior Normalized Jerk Score (APJ score), a measure of smoothness of walking calculated with respect to the antero-posterior axis. A low value of the APJ score indicates rigidity, while high values of the index indicate too much fluidity: For instance, people with Parkinson's disease have values above the normal range. Since the considered subjects do not have important illnesses compromising their walking ability, we can say that people with higher values of the APJ score have a better physical condition.

In our sample, the score takes values from 0.1695 to 2.9811, with an average of 0.9003.

To exploit the longitudinal nature of these data, we rely on growth curve models (Hedeker and Gibbons 2006). These models allow to describe individual patterns of the above parameters during the walking test, modeling the change over time, where the 44 straights (LAPs) represent the time. Moreover, the model allows to detect how subject's characteristics influence the test performance. The selection bias due to missing values is taken into account by means of a pattern mixture model (Hedeker and Gibbons 1997).

Let be Y_{ij} the APJ score for the ith subject in the jth LAP, $f_s(t_{ij})$, with $s = 1, \ldots, S$ a set of functions of time, $\mathbf{x}_{ij}$ a vector of time-varying covariates, $\mathbf{w}_i$ a vector of time-constant covariates (e.g. gender), where $i = 1, \ldots, N$ is the subject index, and $j = 1, \ldots, n_i$ the number of completed LAPs for the ith subject. The general growth curve model for the APJ score as a function of time is:

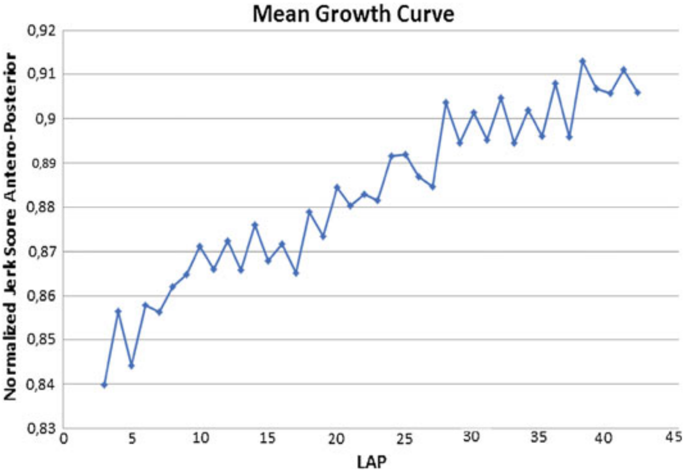


Fig. 1 Normalized APJ score: mean curve

$$Y_{ij} = b_{0i} + \sum_{s=1}^{S} b_{si} f_s(t_{ij}) + \sum_{k=1}^{K} \gamma_k x_{ijk} + \sum_{l=1}^{L} \delta_l w_{il} + \varepsilon_{ij} \tag{1}$$

The model assumes random effects $\mathbf{b} = (b_{0i}, b_{1i}, \ldots, b_{Si})'$ independent from the level 1 errors ε_{ij}. Moreover, we assume normal distributed level 1 errors, i.e., $\varepsilon_{ij} \sim N(0, \sigma^2)$, and multivariate normal random effects, $\mathbf{b} \sim MN(\boldsymbol{\beta}, \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is the covariance matrix of the random effects, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_S)$ is the mean vector.

The average curve is reported in Fig. 1. This curve is obtained fitting model (1) without covariates, considering a parameter for each LAP, namely with the time functions $f_s(t_{ij}) = 1$ if $j = s$ and zero otherwise, and constant slopes $\mathbf{b} = \boldsymbol{\beta}$. The average curve shows a linear increasing trend; thus, on average, the APJ score increases constantly during the walking test.

However, each subject in the sample has a different pattern. As an example, Fig. 2 shows the APJ score pattern for 12 subjects. For some subjects, like number 482, the APJ score is constant during the test, while for others, like subject number 1806, the APJ score increases, with highs and lows. The curve of subject number 246 is interrupted, since this subject stops before the end of the test. To take into account the variability between individual patterns, we specify a linear growth curve model with random slopes.

We fit a linear random slope model, thus allowing a different initial value of the APJ score and a different rate of growth for each subject:

$$Y_{ij} = b_{0i} + b_{1j} LAP_{ij} + \varepsilon_{ij} \tag{2}$$

Model estimates are obtained by means of the mixed procedure of Stata (2015), Rabe-Hesketh and Skrondal (2012). Model results are reported in the first column

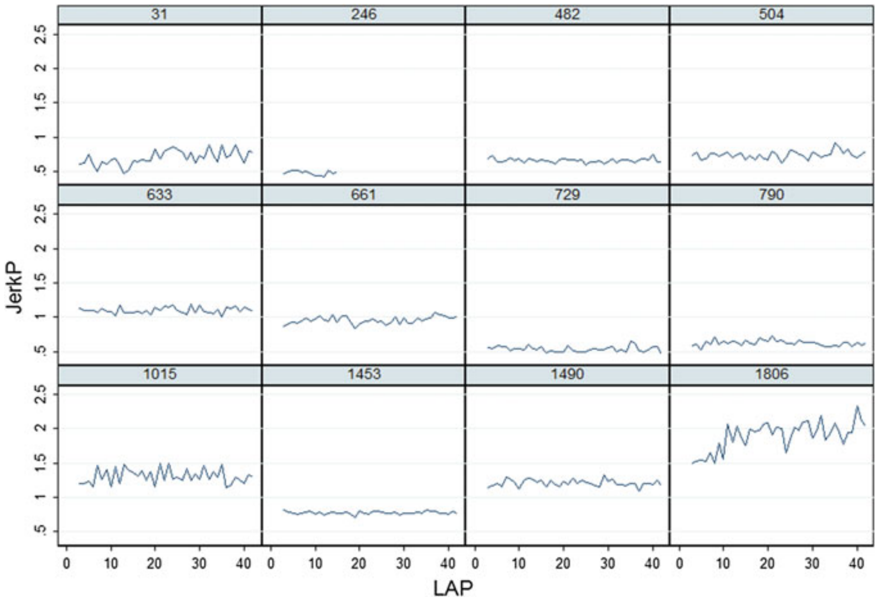


Fig. 2 Pattern of the normalized APJ Score for 12 subjects

of Table 2. The average level of the APJ score at the first considered LAP (the third) is given by $\hat{\beta}_0 = 0.8522$. The average rate of growth is positive, $\hat{\beta}_1 = 0.0015$; thus, subjects tend to become smoother as time goes by.

However, there is a lot of variability around these average values. Indeed, the 95% coverage interval for the random intercepts is (0.3058, 1.3986), while the random slopes range in the interval (−0.0032, 0.0062); thus, the rate of growth could also be null or negative for some subjects. This is coherent with the trends shown in Figs. 2 and 3.

We continue the analysis adding subject characteristics, i.e., time-constant covariates at level two, with the aim to explain intra-subject variability and study how the curves modifies on the basis of these characteristics.

First, we use pretest variables, such as age, height, gender, and the result of Mini-Mental test. Only three out of these covariates turn out to be significant: age at the interview, height, and declared physical activity, an ordinal variable, measured on a tree-point scale (1 = low, 2 = medium, 3 = high).

In order to simplify the interpretation of the intercept, we centered age and height at the minimum value observed among sampled subjects; thus, age = 0 corresponds to 35 years, and height = 0 corresponds to 135 m.

Model results are reported in the third column (named ModCov 1) of Table 2. According to the model, the initial value depends on observed and unobserved individual characteristics. In particular, different values of age, declared physical activity, and height shift the average line downward or upward, according to the sign of

Table 2 Growth curve models of the normalized Antero-Posterior Jerk score

Variables	Random slope	ModCov 1	ModCov 2
constant (β_0)	0.8522 ***	0.7961 ***	0.0996
LAP (β_1)	0.0015 ***	0.0015 ***	0.0012 ***
Age		−0.0057 ***	−0.0036 ***
Physact		0.0729 **	0.0522 *
Height		0.0043 **	0.0063 ***
HarRatioAPm			0.2690 ***
Var.Coef. of Cadence			0.0104 *
IndSimAPm			−0.0989
SD(IndSimAP)			0.5294 ***
LAPXSD(IndSimAP)			0.0041 ***
Variability of random effects			
SD(υ_0)	0.2788	0.2432	0.2209
SD(υ_1)	0.0024	0.0024	0.0023
SD(ε)	0.0640	0.0640	0.0640
Goodness of fit statics			
AIC	−28719.8	−28700.094	−28766.024
logL	14364.9	14358.047	14396.012

Note: * p-value < 0.05, p-value < 0.01, *** p-value < 0.001

their coefficients: The smoothest subjects with respect of the antero-posterior axis are young, tall, and practice physical activity.

The introduction of these covariates explains about 13% of intercepts variability.

The model assumes different individual rates of change. In order to explain this variability, we considered the interactions between subjects covariates and LAP, but they turned out to be not significant.

In the next step, we add to the model subject averages and standard deviations of some measures collected during the test: Indexes of Symmetry, Harmonic Ratios and Variation Coefficient of Cadence. The repeated measures of such variables are quite constant, within the subject. The average and the standard deviation of these variables are reported in Table 1. Model results are reported in the fourth column (ModCov 2) of Table 2. The individual mean of the Antero-Posterior(AP) Harmonic Ratio has a positive effect, as we expected since this index describes the same characteristic of the response, i.e., smoothness of walking. The individual mean of the Variation Coefficient of Cadence has also a positive effect, which means that less coordinated subject are also more smooth. The individual Standard Deviation of the AP Index of Symmetry has a positive effect, but in this case there is also a positive interaction with LAP. Thus, we can say that subject with a less constant Index of Symmetry tends to be more smooth at the begging and also tends to become smoother during the test. Indeed, the estimated rate of growth is $0.0012 + 0.0041 \times SD(InSimAP)_i$; thus, a subject with a Standard Deviation of the AP Index of Symmetry equal to

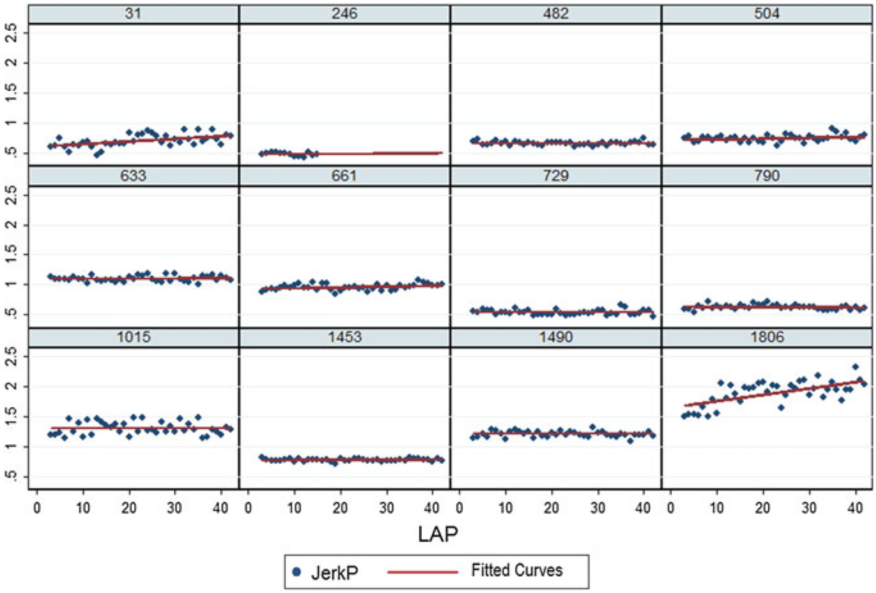


Fig. 3 Random slope model: estimate trend and observed values for 12 subjects

0.0844 (the average SD) has an estimated rate of growth of 0.0015, while a subject with a Standard Deviation of the AP Index of Symmetry equal to 0.0121 (minimum SD) has an estimated rate of growth of 0.0063 m.

The introduction of these covariates explain a further 9% of intercepts variability and about 4% of slopes variability.

As shown by the dropout indicator in Table 1, about 9% of the subjects do not complete the walking test, maybe due to physical problems preventing them to walk 400 m, and we tested whether model covariates are able to take into account the inner differences between the subjects who completed and subjects who did not complete the walking test. To this end, we consider an extension of model (1), namely the pattern mixture model (Hedeker and Gibbons 1997). In particular, we consider two groups: subjects who complete the 44 LAPs and subjects who stop before the end, and we insert a dummy variable which classify subjects in groups on the basis of their missingness pattern. We also insert an interaction term between such variable and the LAP, thus allowing for non costant effect of missing pattern. The parameters of these two added variables can be considered to perform a test for differences between the two groups of subjects. Since we found that these coefficients are not significant, the missing at random assumption holds for the Normalized Antero-Posterior Jerk Score, given the covariates.

Table 3 Logit regression for the probability of fall

Variables	Logit model
Constant	−0.0190
SPS	−0.1877 ***
IndSimAPm	−3.1152 ***
SD(IndSimAP)	5.3344 ***
Goodness of fit static	
LogL	−4902.4977

Note: * p-value < 0.05, p-value < 0.01, *** p-value < 0.001

4 Prediction of Falls

The subjects were monitored for 12 months after the walking test, to record potential falls. About the 15% of the sample (47 subjects) experienced at least one fall during the year after the test.

We want to investigate if the 400-m walk test is a good instrument to predict falls. To this end, we fit a logit model (Long and Freese 2006) for the probability to fall, considering all the pretest variables, the mean value and standard deviation of the variables taken during the test, and the dummy variable identifying subjects who do not complete the test. Model estimates are obtained by means of the logit procedure of Stata. Table 3 reports the selected model, retaining only significant covariates. It is worth to note that not completing the walk test is not informative for future falls. The only covariates who help predicting future falls are the Antero-Posterior Index of Symmetry (mean and standard deviation) and the result of the Short Performance Score (SPS). The coefficient of the Index of Symmetry is negative and, thus, subjects with a symmetric gait have less probability of fall; however, this probability increases with the variability of the Index of Symmetry. The Short Performance Score is a physical test performed before the 400-m walking test, taking values from 1 to 12: the higher the score, the lower the probability to fall. The estimated logit model exploits only few information of that available in the data. We will further analyze data considering Latent Growth Curve Models (Muthén 2004; Bartolucci and Murphy 2015) to cluster subject with respect to the way they perform the test and use this classification to predict falls. Moreover, in this paper, we rely on the APJ score to describe walking patterns; however, this is only one of the many measures collected during the test. We will account for the multivariate nature of response fitting a Multivariate Growth Curve Model (McCallum et al. 1997). This will help to better discriminate subjects with different characteristics in performing the walk test.

References

Bartolucci, F., & Murphy, T. B. (2015). A finite mixture latent trajectory model for modeling ultrarunners' behavior in a 24-hour race. *Journal of Quantitative Analysis in Sports*, *11*(4), 193–203.

Chang, M., Cohen-Mansfield, J., Ferrucci, L., Leveille, S., Volpato, S., de Rekeneire, N., et al. (2004). Incidence of loss of ability to walk 400 meters in functionally limited older population. *Journal of the American Geriatrics Society*, *52*(12), 2094–2098.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, *2*, 64–78.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. USA: Wiley Series in Probability and Statistics.

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using stata* (2nd ed.). College Station TX: Stata Press.

McCallum, R. C., Kim, C., Malarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate multilevel change using multilevel models and latent curve models. *Multivariate Behavioral Research*, *32*(3), 215–253.

Muthén, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences*. Sage Publications.

Newman, A. B., Simonsick, E. M., Naydeck, B. L., Boudreau, R. M., Kritchevsky, S. B., Nevitt, M. C., et al. (2006). Association of long-distance corridor walk performance with mortality, cardiovascular disease, mobility limitation, and disability. *JAMA*, *295*(17), 2018–2026.

Palmerini, L., Mellone, S., Avanzolini, G., Valzania, F., & Chiari, L. (2013). Quantification of motor impairment in Parkinson's disease using an instrumented timed up and go test. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *21*(4), 664–673.

Rabe-Hesketh, S., & Skrondal, A. (2012). Multilevel and longitudinal modeling using stata. In *Continous responses* (Vol. 1). College Station, Texas: Stata Press.

Sayers, S. P., Guralnik, J. M., Newman, A. B., Brach, J. S., & Fielding, R. A. (2006). Concordance and discordance between two measures of lower extremity function: 400 meter self-paced walk and SPPB. *Aging Clinical and Experimental Research*, *18*(2), 100–106.

Stata., (2015). *Multilevel mixed-effects reference manual*. College Station, Texas: Stata Press.

Tian, Q., Resnick, S. M., Ferrucci, L., & Studenski, S. A. (2015). Intra-individual lap time variation of the 400-m walk, an early mobility indicator of executive function decline in high-functioning older adults? *AGE*, *37*, 115.

Tian, Q., Simonsick, E. M., Resnick, S. M., Shardell, M. D., Ferrucci, L., & Studenski, S. A. (2015). Lap time variation and executive function in older adults: The baltimore longitudinal study of aging. *Age Ageing*, *44*(5), 796–800.

Vestergaard, S., Patel, K. V., Bandinelli, S., Ferrucci, L., & Guralnik, J. M. (2009). Characteristics of 400-meter walk test performance and subsequent mortality in older adults. *Rejuvenation Research*, *12*(3), 177–184.

Recurrence Analysis: Method and Applications

Maria Carmela Catone, Paolo Diana and Marisa Faggini

Abstract The awareness that phenomena (social, natural) are for the most part complex and consequently require more realistic models has led to the development of powerful new concepts and tools to detect, analyse, and understand non-stationarity and apparently random behaviour. Almost all existing linear and nonlinear techniques used for the study of time series presume some kind of stationarity, but the application of such tools to non-stationarity and apparently random time series produces misleading results. Recurrence analysis is an advanced technique for nonlinear data analysis used to identify the general structure, non-stationarity, and hidden recurring elements in a time series. Differently from traditional time series techniques that previously assume the nature of the series, the recurrence analysis can be conceived as a diagnostic tool which provides an exploratory analysis identifying the structure of the series. After a general overview of the epistemological and technical underpinnings for the emergent concepts of complexity and nonlinearity, this paper examines the main features of the technique through theoretical examples and a significant review of the main applications.

Keywords Recurrence analysis · Time series · Nonlinearity · Complexity

The original version of this chapter was revised: Author name has been included. The erratum to this chapter is available at https://doi.org/10.1007/978-3-319-55477-8_28

M.C. Catone (✉) · P. Diana
Department of Political, Social and Communication Sciences,
University of Salerno, Fisciano, SA, Italy
e-mail: mcatone@unisa.it

P. Diana
e-mail: diana@unisa.it

M. Faggini
Department of Economic and Statistic Sciences, University of Salerno,
Fisciano, SA, Italy
e-mail: mfaggini@unisa.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_14

1 Introduction

For a long time, science has been considered as a set of activities consisting of explaining, predicting, and verifying phenomena in terms of cause–effect relationships. This perspective of inquiry, mainly based on the verification of general laws and an objective, measurable, predictable idea of reality, underwent an important phase of transformation at the end of the nineteenth century which marked a change for the scientific rationality model in both natural and social sciences. More specifically, in contrast to the predominant idea of science based on perfect determinism, the correct prediction and repeatability of results, and a conception of reality as simple and ordered, a new model of thinking emerged which recognized the nature of science as bounded, probabilistic, pluralistic, and constantly evolving. The change of perspective derives from the recognition of complex phenomena (Eve et al. 1997; Byrne 2013; Byrne and Callaghan 2013), where behaviour is not necessarily determined by proportional relationships of cause and effect. This informs a weaker but more realistic conception of determinism (Bertuglia and Vaio 2003), which can also account for emergent and multidimensional phenomena. The epistemological aspects involved also methodological and technical issues. In the case of the time series analysis, a spectrum of models has been developed, ranging from the classical to modern approaches (Chatfield 2013). While according to the classic approach, the time series can be analysed through decomposition into trend, cycle, and seasonality components, modern methods assume that time series should be conceived as the realization of a stochastic process. In particular, Slutsky, Walker, Yaglom, and Yule first developed the concept of autoregressive (AR) and moving average (MA) models, i.e. linear dynamic models that generate stationary processes (De Gooijer et al. 2006). Next Box, Jenkins, and Reinsel popularized an approach that combines the moving average and the autoregressive approaches, extending it also to non-stationary processes through the autoregressive integrated moving average (ARIMA) models with the argument that such models better represent real data behaviour (Box and Jenkins 1970; Box et al. 2015). Linear models interpret all regular structures in a data set through linear correlations (Kantz and Schreiber 2004) and are usually easy to perform and understood but present limitations. They fail to detect strong asymmetries, irreversibility, and noise amplification in the data, and they are not suitable for identifying nonlinear dependency structures (Tong 1990; Kantz and Schreiber 2004; Giannerini 2012). In particular, the limits of linear methods emerge when they are faced with complex phenomena which require more realistic models in order to detect, analyse, and cope with non-stationarity and apparently random behaviour. The need for describing real phenomena has led to the introduction of new models such as nonlinear ones. In the literature, there are many examples of nonlinear methods used to analyse time series such as the "ARCH-type" models generalized by Bollerslev and the integrated (IGARCH) and fractionally integrated model (FIGARCH) which are based on the assumption that data are nonlinear stochastic functions of their past values (Bollerslev 1986). Both linear and nonlinear techniques are usually based on the assumption that the phenomena investigated follow specific behaviour; in other words, these techniques suppose the characteristics of the series before confirming their presence. Among the nonlinear time series tools is recurrence analysis (RA) which is the object of study for this paper. RA is a technique introduced by Eckmann et al. (1987) designed to identify hidden recurring patterns and structural changes as well as to explore data correlation in the time series. In contrast to some traditional time series techniques that preliminarily suppose the nature of the series, RA does not make any assumptions regarding the statistical distribution and the stationarity of the series (Strozzi et al. 2002). As will be presented in this paper, the technique can be conceived as a diagnostic tool that allows us to recognize the characteristics of the series and in this sense provide an exploratory analysis identifying the structure of the series. Another strength of the technique lies in its diversity of application: RA can be applied to both short and long, high dimensional data sets which represent the new frontiers for the empirical basis of social and natural sciences. RA consists of a graphical analysis in the form of a recurrence plot and a numerical analysis which uses specific quantitative ratios. The paper is structured as follows: Sect. 2 is devoted to exploring the recurrence plot, while Sect. 3 focuses on recurrence quantification; lastly, Sect. 4 provides a significant review of the main RA applications.

2 Recurrence Plot

The recurrence plot (RP) is a graphical method designed to locate the main characteristic of the time series, in terms of recurring patterns, non-stationarity, and structural changes. The plot is based on the reconstruction of time series by substituting each observation X(t) in the original signal with a delayed vector; then, the distances between these vectors are displayed in the plot. More specifically, the plot is built by “putting” the series into a multidimensional space of vectors whose coordinates are the previous and present values of the series. The vectors are constructed according to two parameters: the embedding dimension and the time delay. The embedding dimension expresses the number of vector components and represents the degrees of freedom for the system; it is estimated by the nearest-neighbour method of Kennel et al. (1992) that serves to immerse the system in a vector space where the true distance between points, representative of the vectors, is identified. The idea provided by this method is to choose the neighbouring points of the series that remain near one another after increasing the dimensions. For example, in Fig. 1a, the A_1 and B_1 points in one dimension are still close in two dimension: this means that they are true neighbours; instead, increasing the dimension from one to two (Fig. 1b), the distance between the neighbouring points considerably changes, suggesting that they are false neighbours.

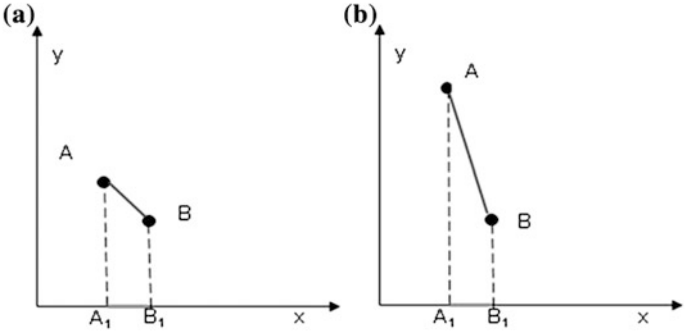


Fig. 1 Nearest-neighbour method: examples of true and false neighbours

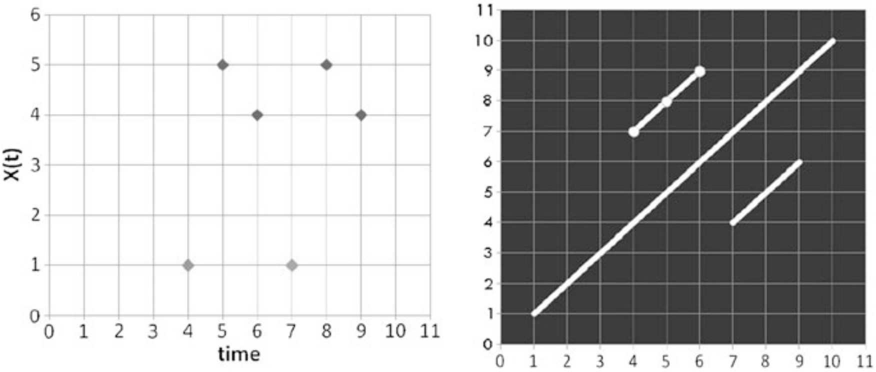


Fig. 2 Examples of time and recurrence plots

Once the embedding dimension has been calculated, the second parameter is time delay[1]; during this time, the relationship between the data is maintained. Next, the Euclidean distances between the vectors are calculated and represented using a colour scheme: short and long distances are usually indicated by light and dark colours, respectively. The main diagonal of the plot, also called "line of identity" (LOI), is white because the distance of each vector with itself equals zero. The presence of light diagonal lines parallel to the LOI indicates a repetition of parts of the series. For example, in Fig. 2, the configuration of the time plot, where the values of the variable on times 4, 5, 6 are repeated on times 7, 8, 9, is represented in the recurrence diagram through two lines which are parallel and symmetrical to the LOI. Light vertical or horizontal lines suggest constant states in which the system persists for some time.

[1] The time delay is calculated by the minimum of the mutual information which is a measure of the link between the following values of the series.

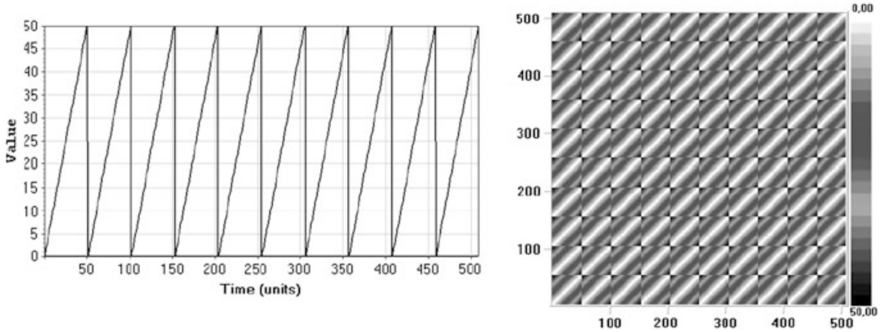


Fig. 3 Time and recurrence plots of a periodic time series

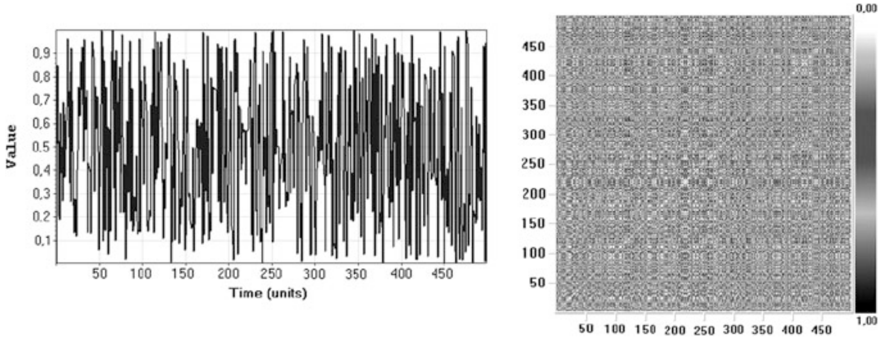


Fig. 4 Time and recurrence plots of a random time series

Some theoretical cases of time series are here proposed in order to better illustrate how the technique works. When considering three typical examples of a periodic, random, and chaotic time series, different configurations of the respective plots emerge. In Fig. 3, the periodic time series RP is made by repeating structures and light diagonals which indicate the periodicity of the signal. In contrast, the random time series RP in Fig. 4 does not show any kind of regular structure, as it is characterized by an absence of light segments parallel to the LOI.

In the case of the chaotic structure,[2] the presence of different size and shade areas and short segments parallel to the LOI suggests a regular series without periodicity (Fig. 5).

By observing the RP, it is possible to gain an overview of the structure of the series and identify its main characteristics. The information provided by the RP, regarding the fundamental nature of the time series, plays a crucial role in allowing

[2] Chaos is aperiodic, deterministic system and sensitive to the initial conditions.

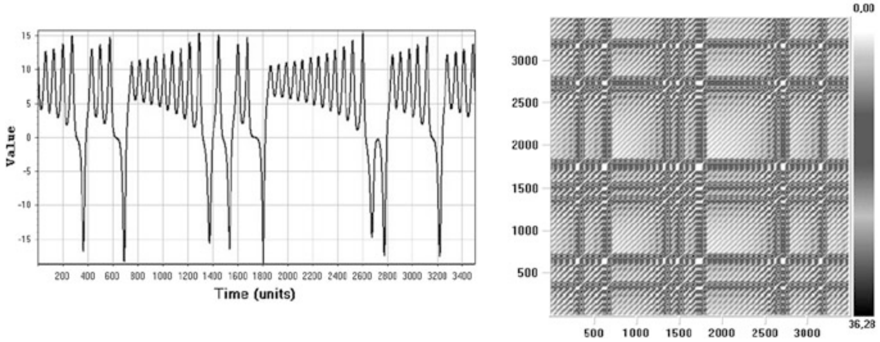


Fig. 5 Time and recurrence plots of a chaotic time series

us to carry out subsequent more specific analysis. Moreover, detailed aspects of the series can be explored by zooming in on selected parts of the plot for closer examination.

3 Recurrence Quantification Analysis

Although RP is a visual tool that allows us to explore the general structure and main characteristics of the series, it is not easy to interpret. Consequently, in order to go beyond the mere visualization of the plot, Zbilut and Webber (1992) developed the Recurrence Quantification Analysis (RQA), a statistical quantification of RP based on a set of measures extracted from the diagonal and vertical segments in a recurrence plot (Webber and Zbilut 2005). The simplest measure of the RQA is the recurrence rate (REC), that is the ratio of all recurrent states to all possible states and expresses the density of recurrence points in the RP. Another RQA measure is determinism (DET), i.e. the ratio of recurrence points that form diagonal structures to all recurrence points; these segments show the existence of deterministic structures, while their absence suggests random structures. The amount of recurrence points which form vertical lines to all recurrence points is expressed by the laminarity (LAM). It is a measure of the graduality of the data variations during time. For example, RQA of the previously described periodic, casual, and chaotic time series provides different REC, DET, and LAM values (Table 1). In particular, REC is prevalent in the periodic series, intermediate in the random series for accidental occurrence and low in the chaotic. DET is maximum in the periodic system where the states are repeated, nil in the random time series, although the REC value is intermediate, and is high in the chaotic series as the trend is replicated, even if without periodicity. Finally, LAM is absent both in the periodic and in random series as the values are not constant, while it is consistent in the chaotic system where successive values do not differ too much from one another.

Table 1 RQA measures

	REC	DET	LAM
Periodic	1.765	100	0
Random	1.029	0	0
Chaotic	0.058	90	26

Another RQA value is the entropy (ENT) which measures the distribution of those line segments that are parallel to the main diagonal and reflects the complexity of the deterministic structure in the system. The trend (TREND) is the regression coefficient of a linear relationship between the density of recurrence points in a line parallel to the LOI and its distance to the LOI; it provides information about the stationarity of the system. The RQA can be performed on the whole series or in specific time interval, obtained splitting the series into epochs. The possibility to detect the RQA for different epochs allows us to develop comparative analysis of different parts of the series, deepening our understanding of their peculiar aspects.

4 Applications and Further Developments

RA has been used in different research areas, showing also a growing trend in its practice during last ten years: from 1987 until now, 1628 references have been identified which include about 1339 empirical studies (source: Nonlinear Dynamics Group, University of Potsdam, 2016). Many applications came from psychology, medicine, and in particular in neuroscience, genomics, and physiology. Also in physics, chemistry, earth science, and astrophysics, engineering, as well as in linguistics, various researches have been developed (Marwan 2003; Webber and Marwan 2015). In recent years, RA has been started to be employed in the social sciences area. An application regards the analysis of daily visit behaviour to an Italian newspaper Website “La Repubblica” registered over three years, from 31 March 2008 to 21 September 2011 (Catone 2013). Different daily time series (number of visits to the Website, time of visit, number of Web pages visited) have been carried out through the RA to identify possible regularities in the use of the Website. The analysis applied to the number of Website visits has provided a recurrence plot characterized by light squared areas around the LOI, in a context of different shade of colour (Fig. 6a). The non-homogeneity of the diagram and the white areas around the diagonal show a non-stationary process. Moreover, in the plot, the white line diagonals emerge, denoting elements of determinism; horizontal and vertical lines indicate signs of laminarity. The presence of squared areas also suggests a certain periodicity. In particular, zooming in on the area around the LOI, a recurrent structure and some periodicity signs are identified. In Fig. 6b, the day-by-day Website visit structure can be detected: the colour of each square point outs that the number of visits during the weekend varies from the visits of the other

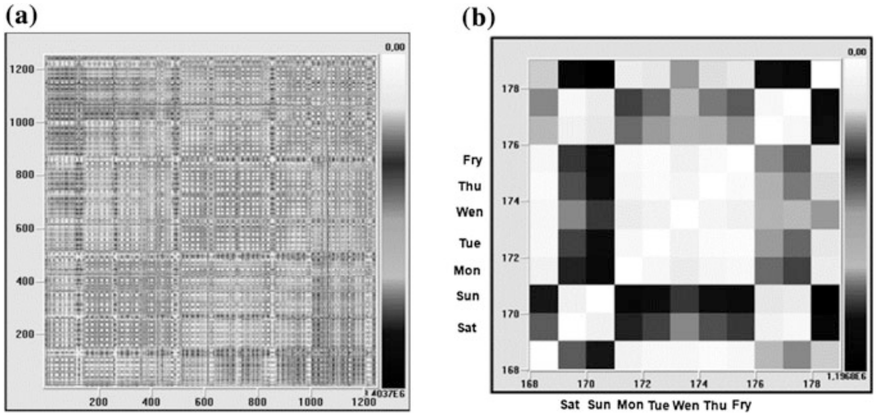


Fig. 6 Recurrence plots of the number of website visits and zoom

days of the week. In this application, RA has been employed to perform an exploratory analysis able to identify the main characteristics of the series.[3]

Other applications came from economics where RA has been adopted in order to find chaos in economics time series and in financial ones (Addo et al. 2013). In this perspective, RA has been used in the study of the macroeconomic time series related to the gross domestic product (GDP) of Japan and UK from 1960 to 1988 (Faggini 2007). The main result of this work has been the identification of chaos in the analysed data, although the shortness of the time series,[4] and the identification of the level of stability in the UK and Japan economies. The RA performed to such data has indicated a nonlinear data relation and a non-stationary process. More specifically, in the Japan and UK time series, the light segments parallel to the LOI in the RP (Fig. 7a, b) and the values provided by RQA have suggested some type of structure that has revealed the presence of a chaotic behaviour. A further evidence of this configuration emerged in the time series has been confirmed by comparing the RP of the original time series (Fig. 7a, b) and its shuffling that did not show any kind of structure, as sign of random system[5] (Fig. 7c).

In other studies, RP and RQA methods have been used also to detect, respectively, a bubble regime and to find the initial bubble time and chaos in a stock market (Neacşu and Todoni 2014). In the social sciences, also other phenomena

[3] A more deeper analysis has also allowed to discover some nonlinear elements and to carry out a short-term prediction [20].

[4] In this research, the macroeconomic time series were already analysed by Frank et al. (1988) with traditional tests for chaos. The analysis carried out through the RA provided different conclusions from the previous research.

[5] The comparison between the original time series and its shuffling is an usual practice in order to confirm a possible regular structure of the original series.

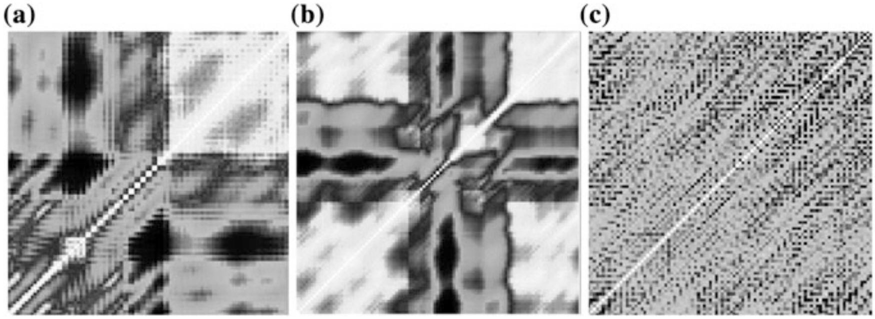


Fig. 7 RP of Japan (**a**), UK (**b**) GDP and Japan shuffled (**c**) GDP time series

have been analysed through RA such as the unemployment rate: in particular, in the series of dynamics of output and unemployment, RQA has allowed the researchers to reveal period characterized by a degree of predictability, with periods of dynamic discontinuity corresponding to the large recessions, like the mid-1970s or the 2009–2010 recessions (Caraiani and Haven 2013); another research deals with the evolution of unemployment transition over time (Chen 2011). Other applications concern the analysis of social interactions (Fusaroli et al. 2014) and the analysis of work motivation and employee well-being (Ceja and Navarro 2011). Although RA can be adopted in short time series, the developments of the technique move towards large data sets which include the analysis of the relation between different time series. In this regard, the new frontiers of RA are the cross-analysis and the joint analysis, i.e. the use of two or more time series, and the recurrence network, a nonlinear approach based on the mix of recurrence analysis and complex network (Iwayama et al. 2013).

5 Conclusions

Recurrence analysis is as an alternative framework for time series analysis as, differently from other techniques that preliminarily suppose the nature of the series, it does not make any assumptions regarding the statistical distribution of the series. This method can be adopted in an exploratory level of data analysis because it provides information on the fundamental structure of the time series. According to the results emerged from RA, the series can be later performed through other subsequent more specific analysis. In social sciences, RA is living an embryonic stage. In our opinion, the technique could respond to new knowledge challenges that aim at identifying and understanding recurrent and latent structures and forms of self-organization in complex social systems (Eve et al. 1997; Castellani and Hafferty 2009; Byrne and Callaghan 2013). The detection and the study of recurrent structures could also foster the developments of social sciences in a predictive perspective.

References

Addo, P. M., Billio, M., & Guegan, D. (2013). Nonlinear dynamics and recurrence plots for detecting financial crisis. *The North American Journal of Economics and Finance, 26,* 416–435.

Bertuglia, C. S., & Vaio, F. (2003). *Non linearità, caos, complessità. le dinamiche dei sistemi naturali e sociali*. Torino: Bollati Boringhieri.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31*(3), 307–327.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. San Francisco: Holden Day.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. Wiley.

Byrne, D. (2013). Evaluating complex social interventions in a complex world. *Evaluation, 19*(3), 217–228.

Byrne, D., & Callaghan, G. (2013). *Complexity theory and the social sciences: The state of the art*. Abingdon: Routledge.

Castellani, B., & Hafferty, F. W. (2009). *Sociology and complexity science: A new field of inquiry*. Berlin: Springer Science and Business Media.

Caraiani, P., & Haven, E. (2013). The role of recurrence plots in characterizing the output-unemployment relationship: An analysis. *PloS One*, *8*(2).

Catone, M. C. (2013). Chaos and non-linear tools in website visits. In T. Gilbert, M. Kirkiolionis & G. Nicolis (Eds.), *Proceedings of the European Conference on Complex Systems 2012* (pp. 87–91). Springer.

Ceja, L., & Navarro, J. (2011). Dynamic patterns of flow in the workplace: Characterizing within individual variability using a complexity science approach. *Journal of Organizational Behavior, 32*(4), 627–651.

Chatfield, C. (2013). *The analysis of time series: An introduction*. Boca Raton: CRC Press.

Chen, W. S. (2011). Use of recurrence plot and recurrence quantification analysis in Taiwan unemployment rate time series. *Physica A: Statistical Mechanics and Its Applications, 390*(7), 1332–1342.

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting, 22*(3), 443–473.

Eckmann, J. P., Kamphorst, S. O., & Ruelle, D. (1987). Recurrence plots of dynamical systems. *Europhysics Letters, 4*(9), 973–977.

Eve, R. A., Horsfall, S., & Lee, M. E. (Eds.). (1997). *Chaos, complexity, and sociology: Myths, models, and theories*. London: Sage.

Faggini, M. (2007). Visual recurrence analysis: Application to economic time series. In M. Salzano & D. Colader (Eds.), *Complexity hints for economy policy* (pp. 69–92). Springer.

Frank, M., Gencay, R., & Stengos, T. (1988). International chaos? *European Economic Review, 32* (8), 1569–1584.

Fusaroli, R., Konvalinka, I., & Wallot, S. (2014). Analyzing social interactions: The promises and challenges of using cross recurrence quantification analysis. In N. Marwan, M. Riley, A. Giuliani & C. L. Webber Jr. (Eds.) *Translational recurrences* (pp. 137–155). Springer.

Giannerini, S. (2012). The quest for nonlinearity in time series. *Handbook of Statistics: Time Series, 30,* 43–63.

Iwayama, K., Hirata, Y., Suzuki, H., & Aihara, K. (2013). Change-point detection with recurrence networks. *Nonlinear Theory and Its Applications, IEICE, 4*(2), 160–171.

Kantz, H., & Schreiber, T. (2004). *Nonlinear time series analysis*. Cambridge: Cambridge University Press.

Kennel, M. B., Brown, R., & Abarbanel, H. D. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A, 45*(6), 3403–3411.

Marwan, N. (2003). *Encounters with neighbours: Current developments of concepts based on recurrence plots and their applications*. Germany: Postdam.

Neacşu, E. L., & Todoni, M. D. (2014). A way to determine chaotic behaviour in Romanian stock Market. *Review of Economic and Business Studies*, 207–217.

Strozzi, F., Zaldívar, J. M., & Zbilut, J. P. (2002). Application of nonlinear time series analysis techniques to high-frequency currency exchange data. *Physica A: Statistical Mechanics and Its Applications*, 312(3), 520–538.

Tong, H. (1990). *Non-linear time series: A dynamical system approach*. Claredon Press.

Webber, C. L., Jr., & Zbilut, J. P. (2005). Recurrence quantification analysis of nonlinear dynamical systems. *Tutorials in contemporary nonlinear methods for the behavioral sciences*, 26–94.

Webber, C. L., Jr., & Marwan, N. (Eds.). (2015). *Recurrence quantification analysis—Theory and best practices*. Cham: Springer International Publishing.

Zbilut, J. P., & Webber, C. L., Jr. (1992). Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters A, 171*(3–4), 199–203.

Part III
On-Line Data Applications

Big Data and Network Analysis: A Promising Integration for Decision-Making

Giovanni Giuffrida, Simona Gozzo, Francesco Mazzeo Rinaldi and Venera Tomaselli

Abstract In recent years, we have witnessed an extraordinary growth in globally generated data. The automatic extraction of such extraordinary amount of data, together with innovative data mining and predictive analytics techniques, represents an innovative opportunity in supporting decision-making. Thus, the main aim of this paper is to explore the opportunity of integrating Big Data techniques with Network Analysis methods. In particular, our study employs descriptive measurements and clustering methods of Network Analysis in order to define relational structures within a Big Data set. We discuss a Big Data tool that collects and analyses information from user interactions with published news and comments about a case study related to a recent Italian constitutional review bill with important political implications.

Keywords Big Data • Network Analysis • Data mining • Decision-making

1 Decision-Making by Network Analysis for Big Data

Today well informed decisions are based on large data sets derived from social media, computing machine logs, and many other related sources. The analysis of Big Data (BD) is useful to support the decision-making process since innovative

G. Giuffrida · S. Gozzo · F. Mazzeo Rinaldi · V. Tomaselli (✉)
Department of Political and Social Sciences, University of Catania, Catania, Italy
e-mail: tomavene@unict.it

G. Giuffrida
e-mail: ggiuffrida@dmi.unict.it

S. Gozzo
e-mail: simonagozzo@yahoo.it

F. Mazzeo Rinaldi
e-mail: fmazzeo@unict.it

F. Mazzeo Rinaldi
ABE School, Royal Institute of Techonology, KTH, Stockholm, Sweden

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_15

and devoted technologies today allow us to manage large amounts of structured data from relational databases and unstructured data. Outcome transparency and the weight of the social information play important roles in decision-making processes (Reader and Leris 2014). Similarly, large networks, as a form of BD, have received increasing amount of attention in data science. Thus, analysing and modelling these data to understand network dynamics is important in a wide range of scientific fields and social science disciplines (Prell 2012).

A critical point about BD is, among others, the inaccuracy danger (McFarland and McFarland 2015). Since observational data from social websites are not derived from statistically rigorous designed experiments, they often can hide biases. The sheer size of the collected data can lead to believe to have census data on an overall population, whereas it is a very biased sample containing a misrepresentative mixture of subpopulations (such as user groups active during the measurement.) BD often produces close-to-zero variances on very large test statistics and then very significant statistical results. Before any analyses on large data sets, the authors propose to employ a community detection data segmentation technique to both identify distinct representative populations at distinct observed activity levels or locations and control for major components of observational data biases. To improve final accuracy, we may need to collect more data in order to adjust for the segment skew and perform better analyses within each segment.

Networks depicting social interactions among actors have been analysed for a long time by a theoretical foundation approach for the application of the Network Analysis (NA) methodology (Cronin 2011). Interaction data may be generated by BD techniques in large scale and at low cost providing detailed information about context, content, and meaning of social interactions as reported events or surveyed opinions (de Nooy 2015). As a formal method for the relational and structural analysis of decision networks, NA has been applied to analyse decision patterns and structures in municipal politics (Laumann and Pappi 1976), decision-making, and organizational behaviour (Cross and Parker 2004).

According to NA approach, the decision mechanisms can be examined as social-relational phenomena. We can measure networks structural attributes such as size, density, clustering, openness, stability, reachability, and centrality (Wasserman and Faust 1994). We can also cluster and categorize structural patterns (i.e. hierarchies, collaboration networks, and chains). Furthermore, we can also apply regression models to test for significant relations between decision structures and performances to assert quality of performing decision patterns.

Integration of BD and NA is not very common in the scientific literature. A study about a large-scale analysis of the news media coverage of the 2012 US presidential elections campaign (Sudhahar et al. 2015) proposes a BD approach as automatic corpus linguistics methods matched with NA. The authors propose a method based on information extraction about the key actors and their relationships in the media narrative of the US elections, organized as a semantic network graph. By the means of this methodology, political positions are automatically derived from a very large corpus of online news, generating meaningful networks with directed links whose semantic nature is retained and understood. The approach is innovative to gain insights in the linguistic analysis of texts by extracting relational data.

To respond to a rising interest for understanding and improving actionable analytics-driven decision patterns based on accurate and meaningful analyses, we propose the integration of a BD model with NA as a method to both extract essential information from a vast data set and identify clusters, patterns, and hidden structures within potential social networks. We test the model on a real case study.

As a first step, we implement the BD model that collects information from user interactions with published news and comments about the real case related to a recent Italian constitutional review bill with important political implications. On the second step, we apply NA on the BD set to extract and analyse essential information from the huge data set to detect useful information to support decision-making process from an empirical methodological perspective.

2 Integrating Big Data and *Network* Analysis: Materials and Methods

2.1 The Big Data Model

The audience model here employed is a generic data model that we can easily apply to many other papers as well (Mazzeo Rinaldi et al. 2017). The two main concepts defined in the model are: *user* and *content*. A user browses the website and performs actions on specific contents. A content is an html document published on a newspaper website. An action is a page view generated by a user on a given content.

Given the set of users, the set of contents, and the actions performed by users on contents, important information can be extracted and analysed. This information helps in understanding the features and characteristics of the users and can be used for different purposes. For instance, it is possible to understand the main users' interests and leverage such information for advertising purposes. This is typically a two-step process. In the first step, contents are classified into categories such as 'Business' or 'Automotive' using advanced text mining techniques. These categories can be interpreted as interests. Then, users are associated with the different categories based upon the stream of contents they read.

In general, every single action a reader performs against a content is stored, together with an appropriate timestamp, in the database. By storing every single action a user performs, even moderate size websites tend to generate very large amount of data. The main challenge, from a technical standpoint, is to cope with such large amount of data. Tools need to be properly devised to make sure such data are promptly available for real-time analysis, which are often needed. Finally, we used a tool to extract the entities contained in the text. These entities have an important semantic value and consequently allow advanced targeting based on sophisticated user profiling.

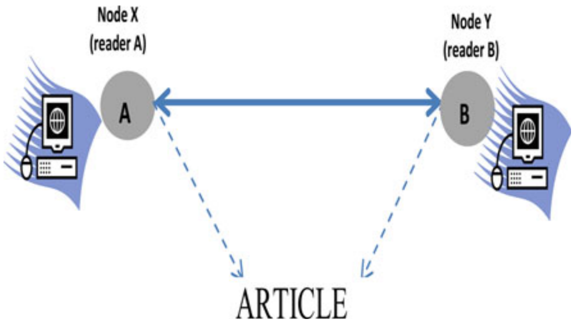


Fig. 1 Structure of data (links)

2.2 *Data Structure for Network Analysis*

This kind of data structure is employed to process data by means of *Gephi* a software for visualization, analysis, and data mining of graphs. The goal is to facilitate the analysis of BD, including the generation of hypotheses, the intuitive discovery of patterns, the isolation of singularities related to relational data sets. Graphical algorithms in *Gephi* are useful to display graphs with many links maximizing the visualization yield compared to the characteristics of the graph (n nodes, n edges, network structure, etc.). This is especially suitable to analyse a huge number of data (nodes and/or edges), visualizing the most relevant relational structures. The best algorithm to obtain a good visualization of many nodes/links is the so-called *Force Atlas 2*. This study is about the analysis of two graphs obtained linking subjects (nodes) who read the same articles (edges) within one of the most important Italian newspaper website (Fig. 1).

When you want to analyse a graph, the nodes are the fundamental survey unit and the links are the unit of analysis. This study analyses links among subjects reading the same news, so each link (line in the graph) represents the relationship between two users (nodes that are connected by the line). Thus if two nodes are connected, that means those readers access the same information (but not necessarily at the same time).

3 Applying the Model to a Real Case

The case we selected to test the model deals with an important constitutional review bill, strongly supported by the Italian prime minister, that lays down provisions to overcome the perfect bicameralism, through the definition of a new differentiated bicameral system. According to this proposal, the vote of confidence in the government will be the prerogative of the chamber of deputies only, while the Senate will be characterized as a representative body of ‘local institutions’. The bill would provide a considerable rationalization of the legislative procedure and the senators will no longer be elected by citizens. Except for constitutional review bills and other constitutional bills, which will still be subject to the current procedures, the chamber of deputies will approve all laws. This bill is still under discussion and a confirmatory referendum will be held at the end of 2016. Within the majority remains a widespread discontent, mainly for choosing to 'steal' the election of senators to the citizens to give it to the regional councillors, not just at the top of popularity in Italy. 'At this point it would be better to abolish the Senate' is one of the most common comments among senators. The constitutional review bill has been widely discussed in the news and generated considerable public interest.

3.1 Results from the Big Data Audience Model

To identify, by hand, 18 articles about the Senate reform, we searched using the query 'riforma Senato', and set a time frame from 1 January 2014 to 31 December 2014. The query returned 47 articles, of which 5 were removed because they contained only pictures and 24 were removed because they were too general and discussed the reform only marginally. We were thus left with 18 articles about the Senate reform, whose effective time frame of publication was from 12 March 2014 until 8 August 2014. All these articles were in the newspaper section about politics. The articles in the section about politics published in the same time frame were in total 1,788. We measured the number of page views and the number of comments of all articles. The 18 specific articles collected together 886,898 page views and 2,461 comments, whereas the 1,788 articles in the politics section collected 32,774,270 page views and 14,108 comments. These statistics have been analysed in order to ascertain the level of interest of the public with respect to the Senate reform, relative to the level of interest of the public with respect to general politics. Finally, the average number of page views per article, the average number of comments per article, and the probability of commenting have been computed. The average number of page views per article indicates the interest of a general reader in the topic. The 18 specific articles have 49,272 average page views, whereas the 1,788 general articles have 18,330 average page views. The average number of comments per articles indicates how many readers are engaged with the topic. The 18 specific articles have 137 average comments, whereas the 1,788 general articles have eight average comments.

Finally, the probability of commenting estimates the likelihood that a comment is written upon reading an article. The 18 specific articles have 0.28% probability of commenting, whereas the 1,788 general articles have 0.04% probability of commenting. Clearly, all these measures indicate that the public is very interested on the topic of the Senate reform.

Next, we analyse the sentiment of each of the 2,461 comments, classifying each comment with a score ranging from -1 to $+1$. Furthermore, we extracted from each of the 18 articles the semantic entities contained in the text of the articles. Both the sentiment analysis and the entity extraction were made using a public API service provided by https://dandelion.eu/. We then generated a spreadsheet containing three columns. The first column contains the extracted entity. The second column contains the average sentiment of all comments. Finally, the third column contains the number of comments upon which the average sentiment has been computed. It was found out that the most popular entities, that is the entities that received the most of the comments, were: ‘Senato’, ‘Governo’, ‘Parlamento’, ‘Senatore’, and ‘Matteo Renzi’. These entities received an average sentiment score of −0.426. Despite the negative sentiment score, this is not to be immediately interpreted as an implication that commenters were against the proposed bill. This is because generally comments in a newspaper tend to be negative (among others: Reis et al. 2015); therefore, the average sentiment score of the top entities must be compared with the average sentiment score of the comments in the general politics section, which was computed to be equal to −0.38. Therefore, it can be inferred that commenters have a slightly negative sentiment (0.04), representing 8% of the standard deviation measured on comments in 0.5, thus insignificant.

We believe that in order to provide more ‘robust’ indications, is necessary, in particular at this stage, an analytical work able to verify and interpret such information. In this phase, some analysts are, for instance, manually analysing a sample of the comments on the 18 articles, to check for correspondence with the sentiment calculated by the software.

3.2 Results from Network Analysis

A further step in this analysis is obtained through NA tools, looking to the co-occurrences of views within the journal website. Links between readers are built removing repetitive and recursive ties. These graphs are obtained analysing the relational structures of two samples:

- The first sample includes edges among all registered readers visualizing the 18 political articles. The graph obtained includes 1.892 nodes and 386.952 edges.
- The second sample includes relations among all registered readers of articles about the Senate reform. The graph on this sample includes 15.550 nodes and 515.739 edges among nodes.

Applying the algorithm *Force Atlas 2* to the website relational data, we obtain a high number of visualizations. It displays the whole graph distinguishing among clusters or groups. This choice makes the structure of relationships graphically more clear than a random solution. *Force Atlas 2* is the most suitable algorithm to display the data. Namely, this algorithm is used to show graphs with high connectivity or with nodes easily reachable, making the graph more readable and offering a clear interpretation of connections among nodes. The application of *Force Altas 2* highlights ‘islands’ of more connected nodes. In these areas of the graph, there are readers who viewed the same news (Fig. 2).

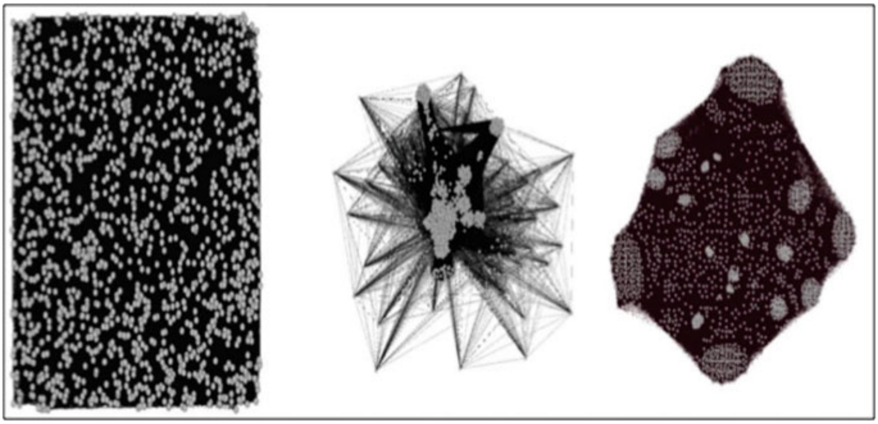


Fig. 2 Layouts from Gephi algorithms of layout: random structure, OpenOrd, and force Atlas2

A first descriptive measure about the structure of network is the graph diameter or the largest geodesic distance: four for the random graph and three for the Senate reform's graph.

This means that no actor is more than three/four steps from any other (close to the small world structure) or there are three/four articles connecting all subjects within the graphs. Although both graphs are compact networks, the second shows a shorter distance among readers and this despite the greater number of nodes. This is not so much due to the number of links (both graphs have a density index of 0.1) but to the structure of networks. Only three articles permit to connect everyone within the graph of Senate reform's readers but the average geodesic (shortest) distance between two nodes is almost one article. This is the number of the most informative articles/items linking the nodes. Another tool to detect the size of connection within the network is the 'eccentricity', a measure of how far each actor is from the furthest other. This procedure shows information about a relative measure (referred to each node), while the diameter is referred to the whole net. The modal eccentricity for both graphs is equal to three, with 1.110 (random articles) and 1.000 (Senate) nodes with this value. Overall, the readers of articles about Senate reform have lower dispersion, focusing on fewer articles more often displayed.

However, the most common NA tool is the 'centrality' measure. There are different measures of centrality. Considering the data structure, the more the node consults articles viewed by other nodes, the more the 'degree' centrality increases. Articles about Senate reform have a higher co-occurrence of consultations. Probably, it is a topic on which focuses the attention of readers, even comparing the data with the graph referred to policy-relevant information. Looking at general political issues, 'degree' distribution shows a higher number of subjects with low degree and a low number of highly 'embedded' nodes. The graph about the Senate reform news shows a less skewed distribution, more concentrated on average values (which confirms the higher average tendency to select the same items). Looking at the diameter, the information is more interesting when you consider that the attention of users is polarized on 3–4 of 18 selected articles (almost three items explain more than 50% of contacts). These results are obtained analysing data by means of both *Gephi* and *Nodexl*, another software useful for NA on big relational data. A second measure is the 'closeness' centrality, applied when an actor is considered important if he/she is relatively close to all other actors. Besides, 'betweenness' centrality is useful to define models based on communication flows and increases the more an actor can control communication flows within the net. Both analysed graphs present high level of degree and low closeness and betweenness. This means that the nets form clusters of readers and that each node has redundant connections (there is a great sharing of information).

The Clauset—Newman—Moore algorithm permits to point out cohesive clusters within the graph, grouping the vertices by means of *Nodexl*. The first graph (with links about political articles) shows intersections among seven collapsed groups, while the second (Senate reform) four clusters (Fig. 3).

Each cluster has a specific relational structure, and it presents specificities if you compare it to the structural characteristics of the whole network. Graph about general political news is (as expected) more heterogeneous. It stands seven different subgroups identified by analysing the relationship dynamic. Besides, there are strong structural differences among groups. First, second, and third groups are the most important, with both high degree (more articles viewed) and betweenness (more brokers) centrality measures. The centrality measure that distinguishes the most relevant subgroups is the betweenness for the first graph and the degree for the second.

What does it mean? NA tools can be useful to drive the decision-making process since they show the collective dynamic among web users. This behaviour cannot be observed by selecting only individual elements. One image we obtain is referred to the diameter and the shortest paths in the graph. This information is about the number or the kind of items that connect all people within the network (more or less spread in the graph or concentrated in specific areas? etc.). Another information is on the identification of subgroups. Each subgroup can be more central or peripheral compared to the whole net. This relational structure could subtend important 'sectorial' articles and users. Looking at our data, the presence of higher

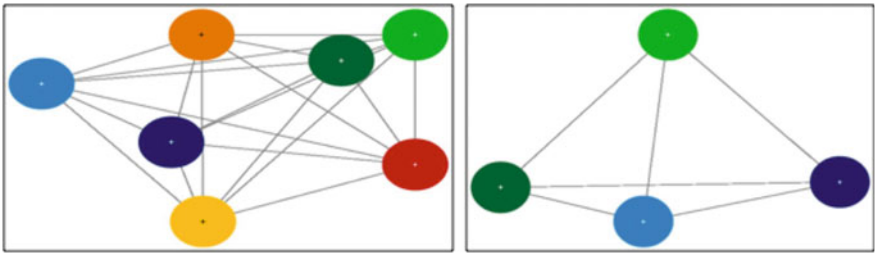


Fig. 3 Collapsed groups on general political news and on Senate reform

betweenness indices for three collapsed groups (first graph) means that there are groups of readers who acquire multiple cross information. A network characterized by subgroups with higher degree centrality (second graph) is more homogeneous with respect to items selected. In this case, few issues convey the interest but there are also the most important readers that can become a reference point for their ability to select quickly significant information.

4 Conclusions

In this paper, we discuss a novel BD analysis approach combining conventional BD analysis techniques with NA. Data processed are real data collected from a very large online Italian newspaper visited by many millions readers each month. In particular, we collected readers' data about a limited hand-picked set of articles discussing a political constitutional reform largely debated in Italy. Readers' interactions with those articles have been collected over a period of one full calendar year, and they include all possible actions performed on such articles such as: reading, sharing, commenting, scrolling. A properly designed BD infrastructure was used to collect such vast amount of data. Such data were then processed in order to make it suitable for being analysed with some NA tools. We present various models for NA which gave us useful insights about the analysed data.

We strongly believe that a wise combination of real-time data collection and BD analytical tools will prove successful in supporting lawmakers and politicians in quickly grasp citizens' sentiment about specific topics. This will be useful in order to shape reforms with the largest possible public acceptance.

In this paper, we prove that technology today is mature enough to provide the right tools to collect and analyse vast amount of data in real time from online publishers and social media. By properly combining such tools, we may support lawmakers with public reaction sentiment measured in a tiny fraction of the time compared with more traditional survey-based techniques.

However, we believe that additional real case and field study have to be carried on. Basically, we need additional support from domain experts in order to shape the suitable data analysis and select the models to process huge amounts of data.

References

Cronin, B. (2011). A window on emergent European social network analysis. *Procedia Social and Behavioral Sciences, 10,* 1–4. doi:10.1016/j.sbspro.2011.01.001.

Cross, R., & Parker, A. (2004). *The hidden power of social networks: Understanding how work really gets done in organizations*. Boston: Harvard Business School Press.

de Nooy, W. (2015). Structure from interactive events. *Big Data & Society,* July-December, 1–4. doi:10.1177/2053951715603732.

Laumann, E. O., & Pappi, F. U. (1976). *Networks of collective action: A perspective on community influence systems*. New York: Academic Press.

Mazzeo Rinaldi, F., Giuffrida, G., & Negrete, T. (2017). Real-time monitoring and evaluation—Emerging news as predictive process using big data based approach. In G. Petersson & J. D. Breul (Eds.), *Cyber society, big data and evaluation* (Vol. 24, pp. 191–214). New Brunswick (NJ): Transaction.

McFarland, D. A., & McFarland, H. R. (2015). Big data and the danger of being precisely inaccurate. *Big Data & Society,* July-December, 1–4. doi:10.1177/2053951715602495.

Prell, C. (2012). *Social network analysis: History, theory & methodology*. London: SAGE Publications Inc.

Reader, S. M., & Leris, I. (2014). What shapes social decision making? (Open peer commentary). *Behavioral and Brain Sciences, 37*(1), 96–97. doi:10.1017/S0140525X13001842.

Reis, J., Benevenuto, F., Olmo, P., Prates, R., Kwak, H., An, J. (2015). Breaking the news: First impressions matter on online news. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media* (pp. 357–366) Oxford University.

Sudhahar, S., Veltri, G. A., Cristianini, N. (2015). Automated analysis of the US presidential elections using big data and network analysis. *Big Data & Society,* January–June, 1–28. doi:10.1177/2053951715572916.

White House Under Attack

Introducing Distributional Semantic Models for the Analysis of US Crisis Communication Strategies

Fabrizio Esposito, Estella Esposito and Pierpaolo Basile

Abstract In the present contribution, the use of distributional semantic models (DSMs) is proposed as a novel approach for the analysis of crisis communication strategies. Temporal Random Indexing (TRI), a specific DSM framework, is employed as computational tool to analyse word meaning change over time. Our resource is represented by the CompWHoB (Computational White House press Briefings) Corpus, a political diachronic corpus collecting the transcripts of the White House Press Briefings from 1993 to 2014. Primary objective of this paper is to demonstrate that TRI can be used in conjunction with Critical Discourse Analysis (CDA) theories as an easily adaptable tool applicable to the analysis of the so-called *crisis communication management* moments where US administration has to cope with risky and serious scenarios.

1 Introduction

Diachronic corpora have recently drawn the attention of the interdisciplinary field of data science, finding application in many research areas ranging from computer science to digital humanities. Providing large amount of data, the investigation and interpretation of these resources has become a pregnant issue today demanding both qualitative and quantitative techniques (Auriemma et al. 2015). In the present contribution, our aim is to propose the use of distributional semantic models (DSMs) in conjunction with critical discourse analysis (CDA) theories as a novel and intuitive approach for the analysis of crisis communication strategies. We employed a specific

F. Esposito (✉) · E. Esposito
University of Napoli Federico II, Napoli , Italy
e-mail: fabrizio.esposito3@unina.it

E. Esposito
e-mail: estel.esposito@studenti.unina.it

P. Basile
University of Bari Aldo Moro, Bari, Italy
e-mail: pierpaolo.basile@uniba.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_16

framework developed by Basile et al. (2014), known as Temporal Random Indexing (TRI), since this model allows to analyse word meaning change over time. The resource on which we operated is represented by the CompWHoB (Computational White House press Briefings) (Corpus Esposito et al. 2015), a political diachronic corpus collecting the transcripts of the White House Press Briefings.

In the first part of the paper, a subcorpus is built following the situational crisis communication theory (SCCT) framework (Coombs 2007). Our analysis approach is then discussed in the second part, describing TRI model, Wodak's CDA theory known as discourse-historical approach (DHA) (Wodak 2001) and the target words selection. Finally, two case studies are shown to demonstrate the suitability of the proposed method.

2 The CompWHoB Corpus

The resource employed in this contribution is the CompWHoB (Computational White House press Briefings) Corpus, a computational implementation of the preexisting specialised WHoB (Corpus Spinzi and Venuti 2013). The CompWHoB, currently being developed at the PRISCA Lab,[1] collects the White House Press Briefings (WHoBs) transcripts extracted from the American Presidency Project website http://www.presidency.ucsb.edu. WHoBs are the daily conferences held by the White House Press Secretary for the news media that are thought to play a crucial role in US administration communication strategies (Kumar 2007); they represent indeed an opportunity for press corps to test White House daily mood, while the administration leverages these meetings to evaluate the course of their communication strategies (Spinzi and Venuti 2013).

The CompWHoB Corpus is linguistically and structurally (Sperberg-McQueen and Burnard 2007) annotated, providing information about the speakers and the date of the event, among other things. At the moment of writing, the corpus spans from 1993 to 2014 [2] and collects more than twenty-five million tokens and more than five thousand briefings.

3 Subcorpus Construction

Since the objective of this paper is to prove TRI robustness and usefulness in the analysis of the *crisis communication management*, we built a subcorpus representative of the these moments of US political life according to the situational crisis communication theory (SCCT) introduced by Coombs. SCCT identifies the key aspects of a specific crisis type predicting the level of reputational threat and the

[1] Department of Physics, University of Napoli Federico II.

[2] The CompWHoB Corpus is planned to be extended to the end of the Obama second term presidency.

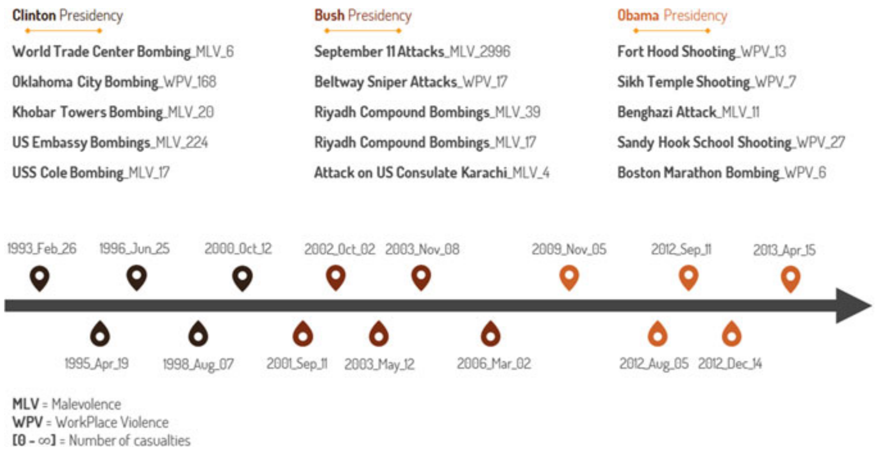

Clinton Presidency
World Trade Center Bombing_MLV_6
Oklahoma City Bombing_WPV_168
Khobar Towers Bombing_MLV_20
US Embassy Bombings_MLV_224
USS Cole Bombing_MLV_17
Bush Presidency
September 11 Attacks_MLV_2996
Beltway Sniper Attacks_WPV_17
Riyadh Compound Bombings_MLV_39
Riyadh Compound Bombings_MLV_17
Attack on US Consulate Karachi_MLV_4
Obama Presidency
Fort Hood Shooting_WPV_13
Sikh Temple Shooting_WPV_7
Benghazi Attack_MLV_11
Sandy Hook School Shooting_WPV_27
Boston Marathon Bombing_WPV_6
1993_Feb_26
1996_Jun_25
2000_Oct_12
2002_Oct_02
2003_Nov_08
2009_Nov_05
2012_Sep_11
2013_Apr_15
1995_Apr_19
1998_Aug_07
2001_Sep_11
2003_May_12
2006_Mar_02
2012_Aug_05
2012_Dec_14
MLV = Malevolence
WPV = WorkPlace Violence
[0 - ∞] = Number of casualties

Fig. 1 Graphic representation of the subcorpus. The events are labelled as either Malevolence or *Workplace Violence*. The arrow represents how the events are distributed along the axis of time

crisis response strategies. In this regard, Coombs defines three crisis clusters based upon attributions of crisis responsibility: the victim cluster, the accidental cluster and the intentional cluster.

Due to the aim of this contribution, we decided to focus on the events belonging to the first cluster, in which the organization is also victim of the crisis and weak responsibility is attributed to it. More in detail, we selected the events defined as *workplace violence*, in which a current or former employee attacks current employees on site, and as *malevolence*, where an external agent causes damage to the organization. Given the nature of our data, we picked out domestic and international acts of terrorism causing at least one fatality related to USA. Hence, we drew upon the Global Terrorism Database,[3] an on-line open-source resource collecting information on terrorist events around the world, in order to select five victim crisis events for each presidency taken into account. Finally, the briefings related to the acts of terrorism were added to the subcorpus using a simple heuristic: WHoBs held on the day of the event and in the following fifteen days were extracted on the basis of the postcrisis phase relevance (Reynolds and Seeger 2005). Figure 1 shows the crisis events collected for each presidency and their distribution across the span of time investigated.

4 Distributional Semantic Models

Distributional semantic models (DSMs) are based on Zelig Harris' *distributional hypothesis* (Harris 1954) assumption stating that the lexical meaning of a word can be inferred directly from its linguistic environment since inherently distributional.

[3] Global Terrorism Database—http://www.start.umd.edu/gtd/.

Hence, words occurring in similar contexts tend to have similar (aspects of) meaning. More formally, DSMs are computational methods that build words meaning as high-dimensional vectors representing the context in which words occur. Thus, placing lexical items in a geometric *word space*, the semantic relatedness between two words can be expressed as the geometric distance between their distributional vectors.

During the last decades, many DSM techniques have been proposed in the literature (Baroni and Lenci 2010; Mikolov et al. 2013). Due to the diachronic characteristics of our corpus, in the present paper, we decided to employ the DSM technique known as Temporal Random Indexing (TRI), since, unlike other classical models, this method is able to manage temporal information making each *word space* comparable to the other.

4.1 Temporal Random Indexing

The Temporal Random Indexing (TRI) is a framework developed by Basile *et al.* that allows to investigate and analyse how words change their meaning over time. Indeed this technique takes into account the temporal information contained in the documents metadata of a corpus and builds several *temporal WordSpaces* for the selected time periods, making them comparable to each other. It follows that word meaning evolution is the result of the comparison between two vectors in two different *temporal WordSpaces*, where words are represented as mathematical points close to each other if showing similar semantic aspects. Hence, the semantic relatedness between two words, computed using the cosine similarity, indicates the word usage variation: the lower the similarity, the greater the word usage variation across two time periods.

In order to perform an in-depth and accurate linguistic analysis, in the present paper we used the open-source TRI software[4] as providing several features for the investigation of word meaning evolution:

- analysis of the neighbourhood of the target word.
- comparison of the neighbourhood of two different target words.
- plot of the semantic shift of the target word over time.
- plot of how relatedness between two target words changes over time.
- clustering of all the vectors belonging to a specific *WordSpace* by using the k-means algorithm.

5 Proposed Analysis Approach

Up to the present, DSMs have been used in a wide range of applications and in various semantic tasks modelling (Bruni et al. 2014). Nonetheless, as also proved by Brigadir et al. (2015), we aim at demonstrating that TRI can be used in con-

[4] Temporal Random Indexing—https://github.com/pippokill/tri.

junction with the discourse-historical approach (DHA) to investigate communication strategies deployed in political contexts. In the next paragraphs, our approach will be discussed in detail.

5.1 Discourse-Historical Approach

Belonging to the field of critical discourse analysis (CDA), Wodak's interdisciplinary and problem-oriented discourse-historical approach (DHA) framework also starts from the assumption that the use of language is a form of social practice. DHA analyses the diachronic change of discursive practices on the basis of the historical sources and sociopolitical context in which they are performed. Since language is considered as a means to reproduce unequal ideologies and power by establishing identity narratives, this approach provides the tools and principles to investigate how discourse strategies are employed to achieve a specific goal. Given the social and linguistic characteristics of our subcorpus, the discursive strategies taken into account to interpret TRI output data are the following:

- Referential: discursive construction of social actors and events through in-group and out-group categorization.
- Predication: discursive qualification of social actors and events by labelling them positively or negatively.
- Argumentation: justification of positive or negative attributions through topoi (e.g. topos of threat, burdening, reality, etc.).
- Perspectivation: framing or positioning of the speaker's point of view through reporting, assumptions and narration of events.
- Intensification/Mitigation: the modification of the epistemic importance of an issue or a proposition by intensifying or mitigating the force of utterances.

5.2 Target Words Selection

Bearing in mind the main objective of our work, we developed a three-step process in order to identify target words representative of the crisis *victim cluster*. The process develops as follows:

1. *A priori* word selection
2. CrisisLex integration
3. FrameNet integration

Following the method proposed by Brigadir *et al.*, 1. represents the starting exploratory step of our process, where a small set of target words is defined and selected *a priori* on the basis of the *crisis* topic and of social context events. This first step is performed after running the TRI on the subcorpus, finding the words changing their meaning the most between two *WordSpaces*. In 2., we integrated our starting set of words with a selection of lexemes drawn from the CrisisLex repository (Olteanu et al. 2014), an automatically generated and human-curated list of terms frequently related to disasters. Due to the different nature of the crisis events, we manually picked out the lexemes considered related to the *victim cluster* events. Finally, in 3. We used the English lexical database FrameNet[5] Baker et al. (1998) to further integrate our lexicon with *lexical units*, namely lemmas evoking a specific semantic frame (Fillmore 1976).

6 Case Studies Analysis

In the following sections, we demonstrate how TRI can be used in conjunction with DHA for the analysis of US crisis communication strategies. Two case studies are provided in order to prove the robustness and suitability of our approach.

6.1 Violence Case Study

After having identified three time periods representing each presidency (Clinton, Bush, Obama), we performed the exploratory step using cosine similarity in order to learn the linguistic terms varying their meaning the most between two *WordSpaces*. From the comparison table generated by the TRI, we selected the word *Violence* as target word since considered a *crisis* context-related term. We focused on Bush and Obama presidencies given the high-usage variation of the *Violence* lexeme.

As shown in Fig. 2, in the Bush *WordSpace* the nouns *Iraq*, *Terrorist*, *World* and *Cooperation* are among the closest terms semantically related to the target word. After further verifying the surrounding nodes semantics, we moved to the identification of the discursive strategies employed by US administration. The *Referential* strategy seems to emerge from the use of the words *Iraq* and *Terrorist*, since constructing an out-group that implicitly categorizes these two entities. At the same time, the semantic relatedness of the terms *World* and *Cooperation* suggests also that an in-group is being constructed asking for support for the issue at hand. Unlike the Bush *WordSpace*, in the Obama one there is no reference to any foreign entities in relation to the target word investigated. Indeed, the first two closest neighbours of *Violence* are the nouns *America* and *Scourge*. In this case, it can be hypothesized the use of both *Perspectivation* and *Intensification* strategies as the crisis manager frames their point of view aligning the *Violence* issue with their country (i.e. America), and the political importance of the topic is emphasized resorting to the metaphoric use of *Scourge*. Moreover, the semantic relatedness of the neighbours *Legislation*, *Laws* and *Congress*, all belonging to the legislative lexical field, may hint at the use of

[5] FrameNet Project—https://framenet.icsi.berkeley.edu/fndrupal/.

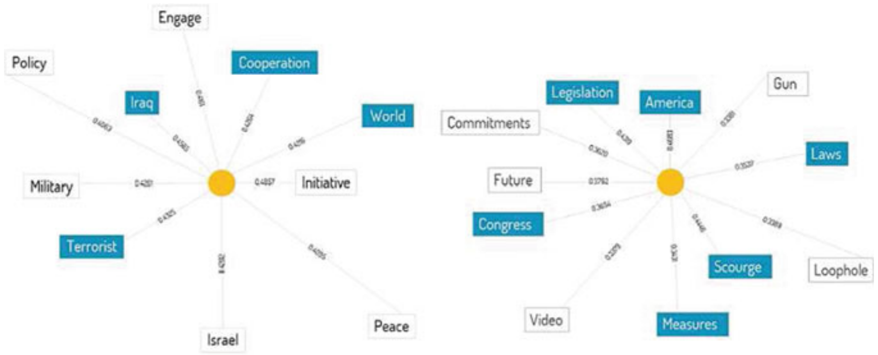


Fig. 2 Neighbourhood graph for *violence* in Bush (on the *left*) and Obama (on the *right*). The *orange node* is the target word. The arcs represent semantic relatedness scores between nodes

the *Argumentation* strategy through the *topos* of reality inferring that some actions should be performed.

6.2 Terrorist Cluster Case Study

Given the nature of the events collected in our subcorpus, we initially picked out the word *Terror* from the CrisisLex repository. Using the FrameNet project, we retrieved the lexical units associated with the target term, deciding to focus our analysis on both the singular and plural forms of the noun *Terrorist*. Employing the TRI, we investigated the semantic neighbourhood of our small cluster in the three *WordSpaces* taken into account. A table was then generated, representing the first ten meaningful neighbour terms related to the issue at hand during each presidency.

In Clinton and Bush *WordSpaces*, the first two closest terms to the *Terrorist* cluster are the pronoun *They* and the demonstrative adjective/pronoun *Those*. As highlighted by Sacks (1992), this adjective/pronoun is often used in rhetorical and political discourse as a means to distinct between the 'self' and 'other'. Hence, the *Referential* strategy can be posited in both the presidencies defining an out-group construction. Furthermore, while in Clinton words such as *Believe*, *Mission* and *Change* may hint at *Perspectivation*, in Bush the use of the *Predication* strategy can be implied. The nouns *Freedom* and *Threat* indeed seem to label the *Terrorists*, identifying them as a collective inferring a threat to the country's values and safety. Unlike the above-said *WordSpaces*, in Obama the first two closest neighbours resulting from TRI analysis are two adjectives: *Preplanned* and *Premeditated*. This stark difference with the previous presidencies can be interpreted as an attempt of establishing an internal logic of the argument, namely an *Argumentation* strategy, where no act of labelling or of out-group creation is carried out. The adjectives *Inexcusable*, *Reevaluating* and the metaphoric term *Grassroots*, all appealing to the semantic field of emotions, also make their appearance among the ten closest neighbours of the *Terrorist* cluster in the Obama *WordSpace*, reinforcing the hypothesis of a new path followed by US administration in dealing with the crisis communication and positing the use of the *Perspectivation* strategy.

7 Conclusions

In the present paper, a novel approach to the analysis of political communication strategies has been proposed. Focusing our attention on those moments of political life where US has been victim of terrorist attacks, two case studies have been investigated. The use of computational techniques such as DSMs in conjunction with CDA theories has been proved to offer a reliable approach and a well-balanced trade-off between qualitative and quantitative methods. Indeed, on the one hand TRI provides the necessary features for an objective semantic exploration, on the other the interdisciplinary and problem-oriented DHA framework offers a more comprehensive understanding of the phenomena under analysis.

The novelty proposed in this work allows not only to gain precious insights about word meaning variation over time but also to comprehend the underlying discursive strategies in use, since rooted in a precise historical and sociocultural context. In our opinion, this approach can prove to be a suitable tool of analysis for both social and linguistic scientists.

References

Auriemma, M., Esposito, E., Iadicicco, L., Marrazzo, F., Polimene, A., Punziano, G., et al. (2015). Euroscetticismo a 5 Stelle: Stili comunicativi e online text data nel caso delle elezioni europee 2014. Sociologia della Comunicazione.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998, August). The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics* (Vol. 1, pp. 86-90). Association for Computational Linguistics.

Baroni, M., & Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, *36*(4), 673–721.

Basile, P., Caputo, A., & Semeraro, G. (2014, December). Analysing word meaning over time by exploiting temporal random indexing. In *The First Italian Conference on Computational Linguistics CLiC-it 2014* (p. 38).

Brigadir, I., Greene, D., Cunningham, P., & (2015, July). Analyzing discourse communities with distributional semantic models. In *ACM Web Science 2015 Conference, 28 June–1 July 2015*, Oxford. United Kingdom: ACM.

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research* (JAIR), *49*, (1–47).

Coombs, W. T. (2007). Protecting organization reputations during a crisis: The development and application of situational crisis communication theory. *Corporate reputation review*, *10*(3), 163–176.

Esposito, F., Basile, P., Cutugno, F., & Venuti, M. (2015, December). The CompWHoB Corpus: Computational Construction, Annotation and Linguistic Analysis of the White House Press Briefings Corpus. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015* (p. 120). Accademia University Press.

Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, *280*(1), 20–32.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2–3), 146–162.

Kumar, M. J. (2007). Managing the president's message. The White House communication operations Baltimore.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014, June). CrisisLex: A lexicon for collecting and filtering microblogged communications in crises. In *ICWSM*.

Reynolds, B., & Seeger, W. M. (2005). Crisis and emergency risk communication as an integrative model. *Journal of health communication*, *10*(1), 43–55.

Sacks, H., (1992). *Lectures on conversations* (Vols. 1 and 2). Oxford: Blackwell.

Sperberg-McQueen, C. M., & Burnard, L. (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange.

Spinzi, C., & Venuti, M. (2013). *Tracking the change in an institutional genre: A diachronic corpus-based study of white house press briefings*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Wodak, R. (2001). The discourse-historical approach. *Methods of critical discourse analysis* (pp. 63–94).

#Theterrormood: Studying the World Mood After the Terror Attacks on Paris and Bruxelles

Rosanna Cataldo, Roberto Galasso, Maria Gabriella Grassia and Marino Marina

Abstract The use of social media has become an increasingly significant phenomenon in contemporary society due to the huge and rapid advances in information technology. People are using social media on a daily basis to communicate their opinions with each other about a wide variety of subjects and general events. Social media communications include Facebook, Twitter, and many others. Twitter is one of the most widely used social media sites and has become an important tool for the assessment of public opinion on various different issues. Recently, several approaches for the evaluation of Twitter messages have been developed, identifying the relationships between words and sentiments associated with relevant keywords or hashtags. In this work, through Twitter, we examine people's reactions to two tragic international events, namely the Paris and Bruxelles terror attacks. Specifically, we have collected comments on Twitter of users from various countries after the attacks. The data were collected using the "twitteR" package in the R programming language; all tweets that contained hashtags such as #notinmyname, #Paris, #PrayForParis, #PrayForTheWorld, #PrayForFrance, and #JeSuisParis from November 27 to December 4, 2015, and all tweets that contained hashtags such as #notinmyname, #PrayForBruxelles, #PrayForBelgium, #Bruxelles, and #PrayForTheWorld from April 5 to 13, 2016, were considered. The textual information was analyzed through techniques of text mining and network analysis in order to detect some important structures of people's communications, so understanding their mood from these threads. Using some R packages, the data were cleaned and analyzed, to classify the tweets into different types of emotion.

R. Cataldo (✉) · R. Galasso · M.G. Grassia · M. Marina
Università Federico II (Napoli), Naples, Italy
e-mail: rosanna.cataldo2@unina.it

R. Galasso
e-mail: roberto.galasso@unina.it

M.G. Grassia
e-mail: mgrassia@unina.it

M. Marina
e-mail: mari@unina.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_17

Keywords Network analysis • Text mining • Terrorism • Twitter

1 Terrorism and Social Media

On the evening of November 13, 2015, the French capital was subjected to a terrorist attack. The attack generated a fervent of activity on social media, to a great extent due to the presence of many people from other countries, a fact which served to create more empathy and compassion among people all over the world. In the hours after the attack, some Parisians started using social media, in particular the Twitter hashtag #PorteOuverte (French for "#OpenDoor"), to offer overnight shelter to strangers stranded by the attacks. The hashtag trended worldwide. A modified version of the International Peace Symbol by London-based French graphic designer Jean Jullien, in which the center fork was modified to resemble the Eiffel Tower, was also extensively disseminated. The symbol was widely shared with the hashtags #notinmyname, #PeaceForParis, #PrayForParis, #PrayForFrance, and #JeSuisParis. Facebook encouraged users to temporarily overlay a transparent image of the French flag to "support France and the people of Paris." Google attached a black ribbon to the bottom of their page "in memory of the victims of the Paris attacks." Skype and other Web sites allowed users to make free calls to France to enable them to connect and communicate with loved ones or relatives/friends in order to be reassured about their safety.

The same ISIS cell behind the Paris attacks launched twin attacks on Bruxelles on March 22, 2016, killing at least 32. Again, as was the case in relation to the Paris attacks, this outrage generated considerable activity on social media.

Social media users around the world showed their support for Bruxelles. As news emerged that dozens of people had been killed or injured in the bombings at Zaventem Airport and at a metro station, Facebook's Safety Check function appeared on the site. This allows users to tell people they are safe and their friends and relatives to check on their loved ones, identifying them as safe. Meanwhile, on Twitter, a number of hashtags started to trend, including #prayforBruxelles and #prayforbelgium. Others posted messages of solidarity with #JeSuisBruxelles, or "I am Bruxelles," a reference to a similar outpouring of emotion associated with #JeSuisCharlie after the Charlie Hebdo attack in January 2015. Some people went beyond merely expressing their solidarity, inviting stranded people into their homes using the #OpenHouse hashtag. The most widely shared image was a cartoon of a French flag comforting a crying Belgian flag with the words "13 novembre… 22 mars…" written underneath. The French cartoonist Jean Plantureux, who goes by the name Plantu, drew this emotional cartoon for the French newspaper Le Monde. A crying person draped in a French flag hugs a crying person with a Belgian flag, suggesting solidarity between the two countries. The dates beneath each figure signify the 13 November Paris attacks and the 22 March Bruxelles attacks.

However, unlike in Paris, accusations of hypocrisy have surrounded the reaction on social media to the bomb attacks in Bruxelles, which many have contrasted with the response to similar terror attacks around the world. The social media reaction to these attacks has been criticized as disproportionate compared to the reaction to those in, for example, Ankara and Istanbul, with many pointing out the significant differences in the responses.

2 Methodology and Approaches

In the literature, there is no standard method for mining and analyzing social media data. Here, an open-source approach for text mining, namely social network analysis, using a set of R packages (Feinerer 2014; Liu 2010) for mining Twitter data and network analysis, is presented.

2.1 Text Mining

Text mining is an automated process for detecting and revealing new, uncovered knowledge and interrelationships and patterns in unstructured textual data resources. Text mining targets undiscovered knowledge in huge amounts of text. In contrast, search engines and information retrieval (IR) systems have specific search targets such as search queries or keywords and return related documents (Gupta and Lehal 2009). This research field utilizes data mining algorithms, such as classification, clustering and association rules, in order to discover and explore new information and relationships in textual sources. It is an interdisciplinary research field combining information retrieval, data mining, machine learning, statistics, and computational linguistics (Gupta and Lehal 2009). First, a set of unstructured text documents is collected. Then, a preprocessing of the documents is performed to remove noise and commonly used words, stop words, and stemming. This process produces a structured representation of the documents known as a term–document matrix, in which every column represents a document and every row represents a term occurrence throughout the document. The final step is to apply data mining techniques, such as clustering, classification, and association rules, to discover term associations and patterns in the text and then, finally, to visualize these patterns using tools such as a word-cloud or tag-cloud (Feldman and Sanger 2006).

2.2 *Network Analysis*

Network analysis is an academic field which studies complex networks such as telecommunication networks, computer networks, biological networks, cognitive and semantic networks, and social networks, considering distinct elements or actors, represented by nodes (or vertices), and the connections between the elements or actors, defined as links (or edges). The field draws on theories and methods including graph theory from mathematics, statistical mechanics from physics, data mining and information visualization from computer science, inferential modeling from statistics, and social structure theory from sociology. The United States National Research Council defines network analysis as “the study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena.”

Network text analysis is a method for encoding the relationships between words in a text and constructing a network of the linked words (Popping 2000).

3 Data and Analysis

This study is based on two datasets:

1. a corpus of 9,660 tweets published by Twitter users in the period from November 27 to December 4, 2015; and
2. a corpus of 15,630 tweets published by Twitter users in the period from 5 April to 13 April 2016.

Each individual tweet in our Twitter collection is normalized and parsed before processing in accordance with the following procedure:

1. The separation of individual terms on white-space boundaries;
2. The removal of all non-alphanumeric characters from the terms, e.g., commas and dashes;
3. The conversion to lower case of all remaining characters;
4. The removal of standard stop words, including highly common verb-forms; and
5. The porter stemming of all remaining terms in the tweet.

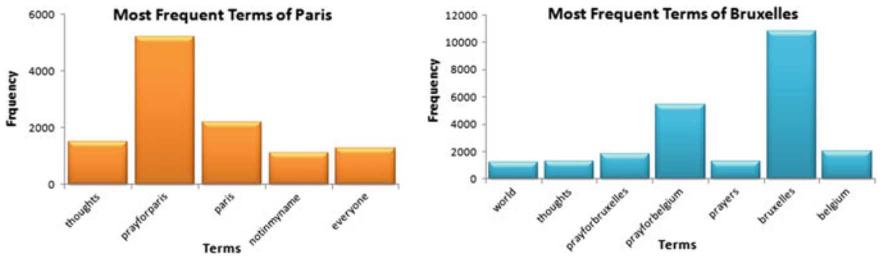


Fig. 1 Plot of the most frequent terms relating to the Paris and Bruxelles attacks

TwitteR, an R package, was used to access the Twitter data. The package enables an authentication of and access to Twitter messages by using keyword search queries (Danneman and Heimann 2014). After the data have been obtained, the extraction of the tweet texts and their cleaning is performed by using the tm package (Feinerer 2014). The output obtained is a structured representation of the tweet texts, the tweet-term matrix. This structured representation can be used to perform text mining using the tm package.

In our study, one of the outputs was a plot of the most frequent terms relating to the two events (Fig. 1).

As regards the dataset of Paris, there are 4,302 words, most having a very low frequency (<500).

The most frequent words are those represented in the chart with a frequency of higher than 1,000, namely “prayforparis,” “paris,” “everyone,” “thoughts,” and “notinmyname.”

Instead, as regards the dataset relating to Bruxelles, there are 6,201 terms used by Twitter users. In this case, the terms with a high frequency (>1,000) are as follows: “bruxelles,” “prayforbelgium,” “prayforbruxelles,” and “belgium,” As we can see in this case, many tweets refer to the nation rather than to the city, whereas in Paris plot, “france” is not among the terms with a high frequency in relation to the Paris attacks. This might indicate that for Web users Paris is regarded independently from the nation.

Another output is the word-cloud representation of the tweets. The clouds are shown in Fig. 2 for Paris and for Bruxelles. In both word-clouds, we have removed terms with a high frequency, in order to see the usage of other words by Twitter users. The size of each term in the cloud indicates the number of mentions of that term in the tweets, thus reflecting its importance.

Terms such as “prayers,” “world,” “peace,” and “love” are used after both attacks, but, as we can see in the Bruxelles word-cloud, tweeters started to use another term that does not appear in relation to Paris: “prayfortheworld.”

This could mean that the second European attack is seen by Web users as a global threat, also because between the two European attacks there were two other attacks, in Ankara and in Istanbul. In fact, Ankara and Istanbul appear in the Bruxelles word-cloud.

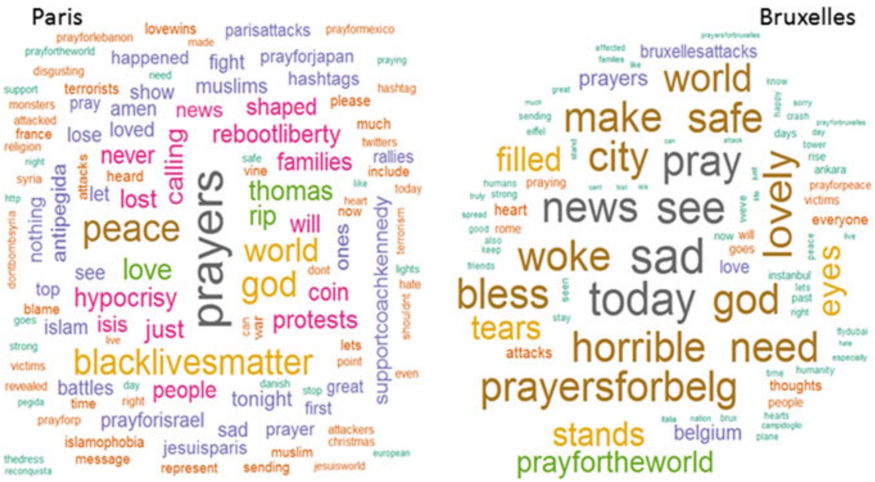


Fig. 2 The word-clouds relating to the Paris and Bruxelles attacks

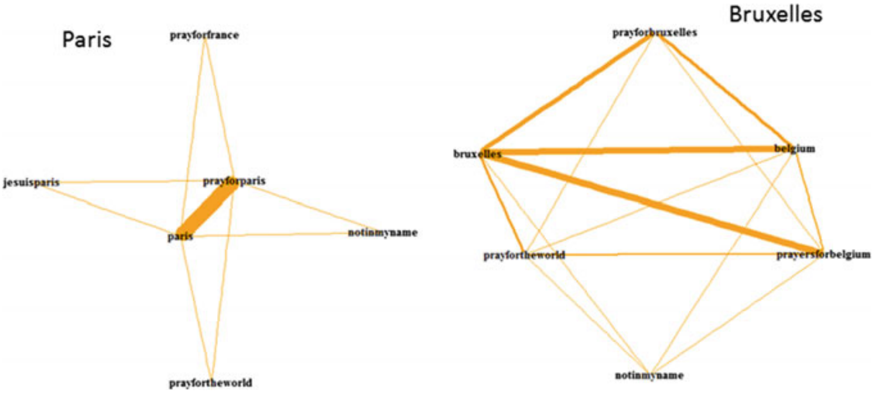


Fig. 3 The hashtag networks relating to the Paris and Bruxelles attacks

Putting it in a general scenario of social networks, the terms can be taken as people and the tweets as groups on Twitter, and the term–document matrix can then be taken as the group membership of people. We have built a network of terms based on their co-occurrence in the same tweets, which relates to a network of people based on their group memberships.

In particular, we have collected only the hashtags for the two datasets. The Paris and Bruxelles networks are represented in Fig. 3. Moreover, in each network, the strength of the connection is represented by the thickness of the relation.

As we can see, in the Paris network, the hashtags connected to all other hashtags are “prayforparis” and “paris,” while there is no co-presence with other hashtags. Moreover, these two hashtags are strongly connected to each other.

Instead, in the Bruxelles network, all hashtags are connected to each other. Only between “prayforbruxelles” and “notinmyname” is there no connection.

This means that Twitter users in their tweets relating to the Bruxelles attacks used all the hashtags simultaneously. In particular “Bruxelles,” “prayforBruxelles,” “belgium,” and “prayersforbelgium” are strongly connected to each other. In contrast, “notinmyname” and “prayfortheworld” are connected with other words, although this connection is not very strong.

4 Conclusions

The coming and diffusion of social media enables people to share opinions and moods, creating a textual corpus of incalculable dimensions updated daily. New statistical techniques involved in analyzing the emotional state and the type of views; through these disciplines, you can understand the mood of a group of individuals, creating a representation of the social emotional state. In this work, as a matter of fact, we have aimed to examine, through Twitter, people’s reactions to two tragic international events: the Paris and Bruxelles terrorist attacks.

The textual information was analyzed through techniques of text mining and network analysis in order to detect some important structures of people’s communications, so understanding their mood from these threads and also going to emphasize such as people’s feelings change over time and in relation to the events.

First of all, we wanted to know which words had been used by Twitter users with a high frequency, and accordingly, in the Bruxelles case, many tweets refer to the nation rather than to the city, while “france,” in the Paris case, is not among the terms with a high frequency. This might indicate that for Web users Paris is regarded independently from France.

Next, word-clouds were designed for both datasets. The analysis of the two word-clouds has led us to understand that the second European attack is seen by the Web users as a global threat, additionally because between the two European attacks, there were two other attacks, in Ankara and in Istanbul.

After the Paris attack, people around the world feel strongly in danger, very exposed, and defenseless in the face of these uncontrollable events. It is also evident from the hashtag networks that there are certain differences. In fact, in the Bruxelles attack appears the word “world”; in the Bruxelles network, it can be noted that “prayfortheworld” is associated with all the other hashtags, going to point out just that terrorist attacks are a worldwide problem and that concerns all of us. All of us are under attack and easily attachable.

References

Danneman, N., Heimann, R. (2014). *Social media mining with R*. Birmingham: Packt Publishing Ltd.

Feinerer, I. (2014). *Introduction to the Tm Package Text Mining in R*.

Feldman, R., & Sanger, J. (2006). *The text mining handbook*. New York: Cambridge University Press. ISBN 978-0-521-83657-9.

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence, 1*(1), 60–76.

Liu, B. (2010). *Sentiment analysis and subjectivity*. In N. Indurkhya & F. J. Damerau (Eds): Handbook of natural language processing (2nd ed.).

Popping, R. (2000). *Computer-assisted text analysis*. SAGE.

Learning Analytics in MOOCs: EMMA Case

Maka Eradze and Kairit Tammets

Abstract The paper overviews the project—European Multiple MOOC Aggregator, EMMA for short, and its learning analytics system with the initial results. xAPI statements are used for designing learning analytics dashboards in order to provide instant feedback for learners and instructors. The paper presents dashboard visualizations and discusses the possibilities of use of EMMA learning analytics dashboard views for sensemaking and reflection of the MOOCs and MOOC experience. It investigates some of the MOOCs in EMMA platform as cases and analyzes the learning designs of those MOOCs. Recommendations of changes to learning designs based on learning analytics data are provided.

1 Introduction

Two major and mainstream innovations that influence the field of technology-enhanced learning are the Massive Online Open Courses (MOOCs) and learning analytics (LA). The first one is regarded as being on the verge of the promise of disruption (Powell and Yuan 2013b) and having a lack of appropriate pedagogical approaches and scenarios.

Daniel (2012) thinks that big provider platforms do not support cMOOCs (a course based on the pedagogical principles of connectivism) and learning interactions are limited. xMOOCs have been criticized for adopting a knowledge transmission model (Powell and Yuan 2013b), and at the same time, cMOOCs provide more space for learner-centered pedagogy. On the other hand, LA is inherently data-driven and quantitative, with lots of click-based logs that different platforms and courses aggregate; the way research questions are hypothesized, operationalized, and results presented is quite simplistic (Reich 2015).

M. Eradze (✉) · K. Tammets
Tallinn University, Narva rd. 25, 10120 Tallinn, Estonia
e-mail: maka.eradze@tlu.ee

K. Tammets
e-mail: kairit.tammets@tlu.ee

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_18

In this article, we intend to show how MOOCs and learning analytics work together on the example of one particular MOOC platform.

2 Learning Analytics—State of the Art

LA is the intersection of many disciplines and operates as an interdisciplinary field; it connects learning sciences, psychology, computer sciences, statistics, and social sciences. LA is a rapidly developing field, and it still is in its infancy. Leveraging its potential to understand learning, predict learners' and teachers' behavior, report, visualize, and act upon the information on different levels for different stakeholders brings some challenges and questions to be answered. It is especially true for analytics in open learning environments such are MOOCs (Chatti et al. 2014). At the same time, learning analytics is a powerful tool for understanding the pedagogies and instructional designs of the MOOCs and investigating the patterns of engagement in informal learning context.

Chatti et al. (2014) views *"LA as a technology-enhanced learning (TEL) research area that focuses on the development of methods for analyzing and detecting patterns within data collected from educational settings, and leverages those methods to support the learning experience."* The authors name some challenges of learning analytics: 1. Big learning analytics that come from heterogeneous sources, 2. open learning environments (MOOCs), with data coming from outside of platform, 3. mobile learning analytics, 4. context modeling, 5. privacy-aware analytics, 6. personalized learning analytics, 7. lifelong learner modeling, 8. learning analytics for open assessment, 9. embedded learning analytics, 10. learning analytics design patterns, and 11. learning analytics evaluation.

Chatti et al. (2014) name four techniques mostly used in LA literature—**Statistics reporting tools** often generate simple statistical information such as average, mean, and standard deviation; **information visualization** (IV)—recognizing the power of visual representations, traditional reports based on tables of data are increasingly being replaced by dashboards that graphically show different performance indicators; **data mining** (DM)—also called Knowledge Discovery in Databases (KDD); and **social network analysis (SNA)**—as social networks become important to support networked learning, tools that enable to manage, visualize, and analyze these networks are gaining popularity.

3 Models and Frameworks and Learning Analytics in MOOCs

It is quite obvious that teaching crowds requires scaling up learning and massive amount of support, which can be implemented through LA. LA in MOOCs can have layers of intertwining challenges and opportunities: First of all, LA is the shortest way to provide personalized feedback to learners and instructors; at the same time, it offers incredible amount of data sources for different stakeholders. All of this combines as a new way how educational research and policy can be planned and implemented. There are some models and frameworks that map learning analytics most important issues, stakeholders, and challenges.

One of the model developed by Chatti et al. (2014) that maps LA between the questions What, Why, How and Who. Building on Buckingham Shum (2012) conceptualization of Learning Analytics levels—micro, meso and macro which corresponds to the following stakeholders' groups—1. Learners and teachers, 2. Institutions 3. Country-wide policies; Drachsler and Kalz (2016) proposes The MOOC learning analytics innovation cycle (MOLAR) framework to view the learning analytics in MOOCs. This is an effort to bring together different domains, objectives, levels of analysis, and processes of LA, and it works at three different levels.

According to Clow (2013), most of the MOOC analytics develop around the formal education context and questions surrounding predictive modeling, that is, problematic in MOOCs. "When learning analytics are most effective, they are an integrated part of a whole system of learner support, which is hard to deliver in a MOOC." Clow believes that *There are two significant points of difference in learning analytics in MOOCs compared to formal education: one qualitative, and one quantitative. The qualitative difference is the rationale behind the course and the aspirations of its designers. In a cMOOC, the designers are explicitly not intending to specify end points before the course starts, so a learner who starts but does not complete may well be seen as a success, depending on the reasons. The quantitative difference is, as the old saw has it, one that is sufficiently large to be a qualitative difference: the rate of dropout is so very much larger in MOOCs. This idea is encapsulated in the funnel of participation.*

According to Chatti et al. (2014), learning increasingly takes place in a networked and open learning environments, and with the increasing popularity of MOOCs, it is especially important. cMOOCs are especially challenging and interesting for their distributed nature that go beyond the platforms.

4 EMMA Project

The European Multiple MOOC Aggregator, called EMMA for short, is a 30-month pilot action supported by European Union. Through large-scale piloting of MOOCs, it aims to showcase the excellence of teaching methodologies in multiple languages. The platform uses automated transcription and translation system and develops innovative learning analytics system based on xAPI specification.

4.1 xAPI

EMMA platform uses xAPI specification for its LA system. The Experience API is a service that allows for statements of experience (typically learning experiences, but could be any experience) to be delivered to and stored securely in a Learning Record Store. The Experience API is dependent on Learning Activity Providers to create and track learning. Learning Activity Provider is a software object that communicates with the LRS to record information about the learning experience. Learning activity is a unit of instruction, experience, or performance that has to be tracked. A statement consists of <Actor (learner)> <verb> <object>, with <result>, in <context> to track an aspect of a learning experience. Several statements can be used to track the whole experience. As xAPI activity statements closely follow the syntax of English (or a logic of syntactic structures of many languages), this data is human readable.

Quoting Verbert, Suthers, and Rosen, Kevan and Ryan (2016) believe that the issue of variety of educational data taxonomies and also the distributed learning events collection challenges can be solved by xAPI: *The xAPI offers a new solution for this issue, using a student-centered approach built on current Web technologies.* xAPI statements are also aligned with constructivist approach (Jonassen and Ronrer-Murphy 1999) and activity theory (Engeström 2001). Kevan and Ryan suggest that simple adoption of the xAPI specification *and configuring a learning content application and LRS repository will not be successful without some other system in place to prescribe meaning to and relationships among activities, experiences, and larger learning goals*. They also believe that xAPI offers a renewed opportunity to research, develop, and explore theories involving learning beyond academia's digital environments.

4.2 EMMA Analytics System—Technical Description

Learning analytics system of EMMA platform consists of following components (see Fig. 1): built-in tracking system that collects the data about users' interactions. The set of users' interactions making up learning experience statements in xAPI format is sent to Learning Record Store (LRS), which is used for storing learning experiences of EMMA courses. Some of the data (e.g., the structure of the MOOC and start and end date of the course) to be visualized by the dashboard was retrieved from EMMA database with the support of Web service. The implementation of visualizations is mainly based on Highcharts charts' framework. Social network analysis graph was developed by using Sigma.js.

The current version of the dashboard is a stand-alone module, which was integrated with EMMA platform, but it could be integrated with different learning environments as far as they will make use of xAPI statements stored in LRS, which does not necessarily have to be Learning Locker and could be any other LRS.

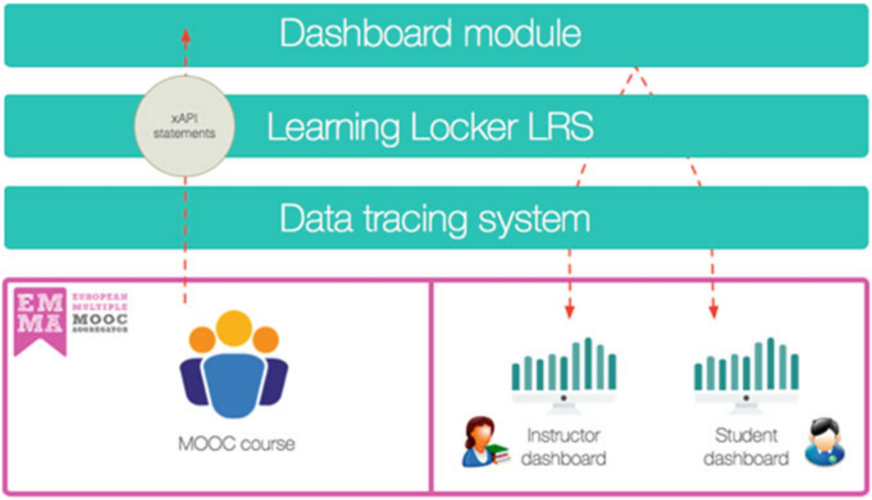


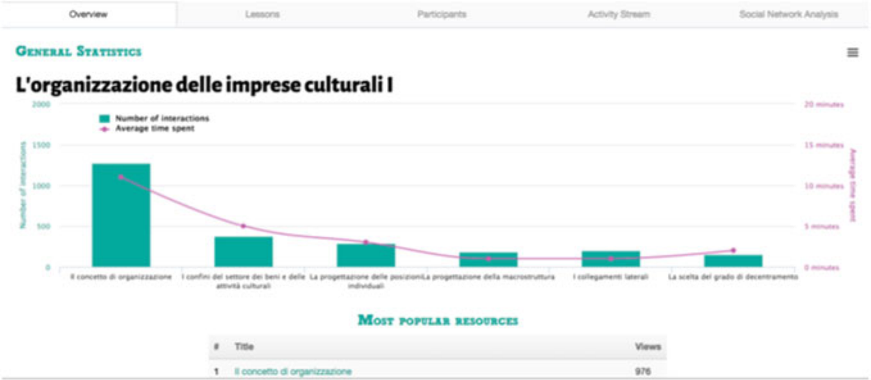
Fig. 1 Technical architecture of the EMMA learning analytics application

4.3 *Dashboards*

Learning analytics dashboard has been developed for the learners and facilitators. Aim of the facilitators' dashboard is to provide them with feedback about their course design and learners' engagement.

Facilitator is provided with the following information in their learning analytics dashboard:

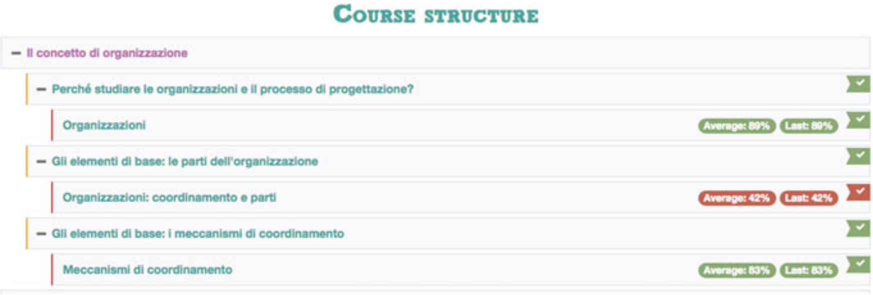
- **Overview of the course activities** against the lessons of the MOOC—interactions in different lessons will be visualized (see Fig. 2).
- **Overview of the different interactions** under the units and lessons (see Fig. 3). This view zooms in the deeper view of the lessons: interactions under different units: assignments, popular learning resources, and discussions between the participants. Each lesson has next dimension, which presents a deeper view or zooms in information. Here, facilitator will be informed how many enrolled students have accessed different units, learning resources, and additional materials under the units;
- **Enrollments** and unenrollments of the participants—daily view of the joined and left participants;
- **Activity stream** of the recent activities—100 latest activities of the MOOC participants will be visualized as stream;
- **Social network analysis**—the represented graph is based on the posts, responses, and comments in conversation module and in assignments (see Fig. 4). Nodes represent the participants, and by hovering the mouse to the node, first and last name of the participants will be visualized.



Most popular resources

#	Title	Views
1	Il concetto di organizzazione	976

Fig. 2 Overview of the MOOC based on the users' interactions


Course structure

− Il concetto di organizzazione

− Perché studiare le organizzazioni e il processo di progettazione? ✓

Organizzazioni Average: 89% Last: 89% ✓

− Gli elementi di base: le parti dell'organizzazione ✓

Organizzazioni: coordinamento e parti Average: 42% Last: 42% ✓

− Gli elementi di base: i meccanismi di coordinamento ✓

Meccanismi di coordinamento Average: 83% Last: 83% ✓

Fig. 3 Learner progress overview page

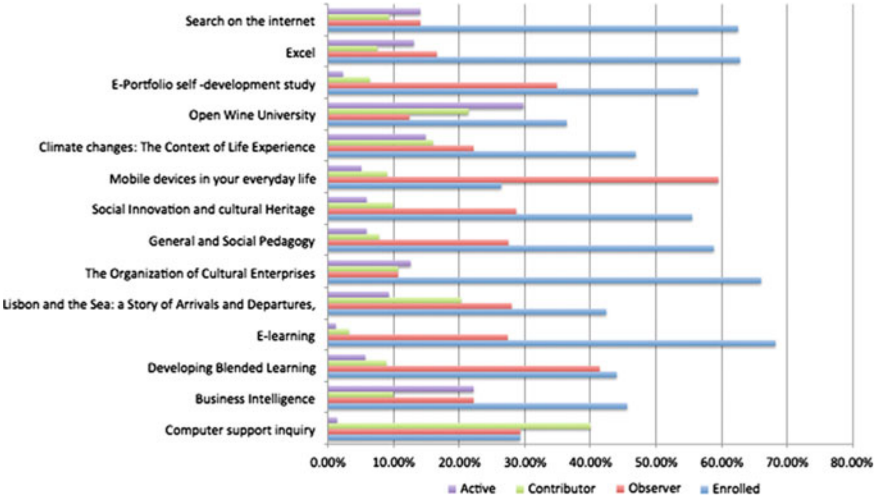


Fig. 4 Clusters of EMMA MOOC participants

Learners' learning analytics dashboard view includes:

- **Progress** compared with different lessons—how many units under each lesson have been accessed by the learner, assignments submitted with what result and participated discussions (see Figure X). These interactions may indicate that learner is passing the MOOC activities and can be used for monitoring progress;
- **Enrollments** and unenrollments of the participants—daily view of the joined and left participants;
- **Activity stream of the recent activities**—100 latest activities of the MOOC participants will be visualized as stream;
- **Social network analysis** based on comments and responses in conversation module of the MOOC (see Fig. 4);
- **Overview of the popular resources** and suggestions to access materials that other participants have accessed, but this certain learner has not yet.

4.4 Data Analysis

To analyze participant's interactions, data from Learning Locker were used. The following interactions were analyzed:

- Learner visited page (lesson, unit, assignment, peer assessment, blogpost);
- Learner commented/replied conversation;
- Learner submitted assignment/peer assessment;
- Duration of different content.

Each MOOC was analyzed separately, and analysis was based on frequencies of interactions. The main functionalities, which may support uptake of other learners' knowledge, are the conversation tool and the blog. The blog is technically not associated with the individual MOOCs, but it is personal and it works across the MOOCs: Therefore, the interactions cannot be used for MOOC analysis. The conversation tool was used for social network analysis.

4.4.1 Clustering the Participants of MOOCs

In the first phase, participants were clustered as enrolled, observers, and contributors. As in the second phase, there were a bit more interactions; the following clustering scheme was used:

- **Enrolled**—participant entered the MOOC up to five times;

- **Observer**—participant entered the MOOC more than five times, but did not interact with the content or other participants;
- **Contributor**—participant contributed with the assignment, comment, or post to the MOOC at least once;
- **Active**—participant contributed with the assignment, comment, or post to the MOOC more than once

Figure 4 indicates that the majority of MOOC participants just enrolled in the courses or observed the content. *Open Wine University* MOOC provided by UoB had more than 50% of the participants who actually contributed to the course. TLU course computer-supported inquiry had more than 40 of the participants who interacted in the platform and Business Intelligence, Lisbon, and the Sea: a Story of Arrivals Departures and Climate Changes. The Context of Life Experiences had more than 30% of participants who interacted in the system.

There might be several reasons why one MOOC has more or less interactions. One of the reasons could be technical (participants cannot find some functionalities), or the course assumes using blogs for interactions, but blogs are not part of the learning analytics analysis (as described earlier). Another reason could be the selected pedagogical design. Some of the courses may not assume that participants contribute to the platform by posting or submitting, but the focus is on reading materials and watching videos.

Students' access materials were also analyzed: There are only few courses, which users' behavior is different compared with others. Most of the courses has similar pattern: 40–60% of users accessed materials less than 10 times (clustered probably as enrolled and some of the observers); 20–40% of the users accessed materials more than 9 times; and 10–20% accessed materials more than 50 times. But course like "computer-supported inquiry" had significantly more learners (40% out of 118 users) who accessed materials more than 50 times.

4.4.2 Social Network Analysis

Social network analysis was based on statements related with commenting, responding, and replying. Figure X illustrates the social network analysis of Open Wine University MOOC. Participants are small nodes. Same picture is visualized on teachers' and participants' dashboards. By moving the mouse on the node, you can see the username related with the node. On this figure, we can see the course facilitator/teacher as the largest node. But nevertheless, there are several interactions between other participants as well. Nodes that are not connected are people who never commented or replied to comments and who could be considered as outsiders of the learning community.

Figure 5 illustrates the course where interactions are even more distributed between other participants. Based on the figure, it is possible to assume that the course had more than one facilitator and that some of the participants were really active commentators. Also, it is possible to see that the direction of the

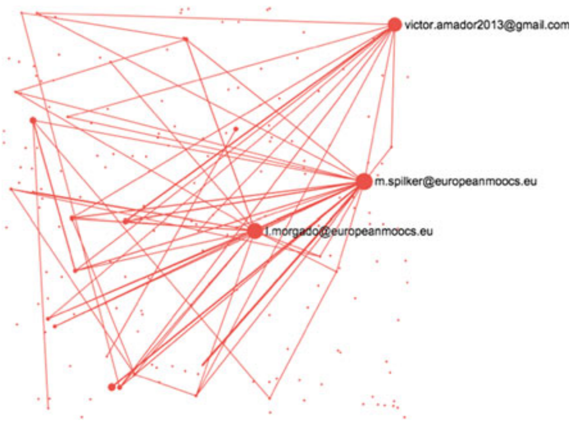

victor.amador2013@gmail.com
m.spilker@europeanmoocs.eu
j.morgado@europeanmoocs.eu

Fig. 5 Social network analysis of the MOOC climate changes: the context of life experience/social network analysis of MOOC computer-assisted inquiry

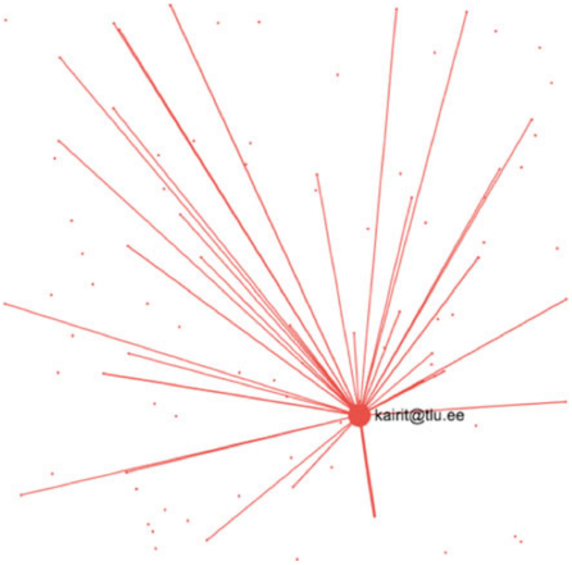


Fig. 6 Social network analysis of the MOOC climate changes: the context of life experience/social network analysis of MOOC computer-assisted inquiry

communication is not only from learner to facilitator, but participants interact with each other quite a lot. Figure 6 describes the course where communication is mainly one-directional. This is one the examples of how the course pedagogical design affects the interactions within the course. In that course, the students were supposed to publish some of their tasks with the conversation tool, so they responded to the facilitator's request with their contribution. It seems that participants did not perceive others' contributions relevant enough, so they did not comment or answer to nearly no comment.

4.4.3 Overview of Participants' Engagement in Different Lessons of MOOCs

For the following analysis, interactions related with time spent, page visits, comments, submits, and responses were analyzed. The aim of the current analysis was to find out how participants engage with the content during the course. Figure X illustrates the amount of interactions and time spent on different lessons based on data of Tallinn University MOOC—*computer-supported inquiry.*

Figure 7 indicates that the amount of interactions and minutes spent on content were slowly decreasing after each lesson. At the same time, Fig. 8 illustrates a different type of progress in the MOOC—*Climate Changes: The Context of Life Experience*. Figure 8 illustrates how in every second lesson there was an increasing trend of interactions and even then people spent more time on the content. Course design may indicate that the participants were supposed to do something more or

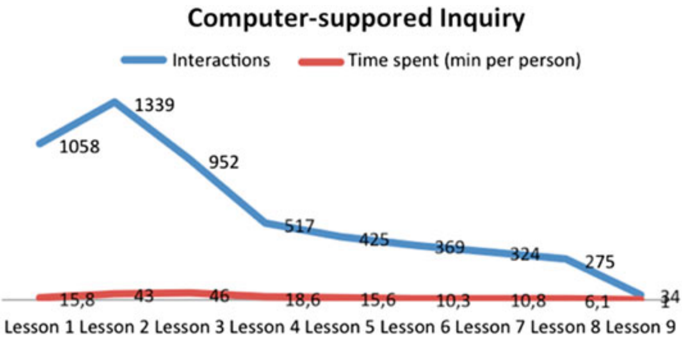


Fig. 7 Interactions and time spent on lessons (computer-supported inquiry MOOC/climate change MOOC)

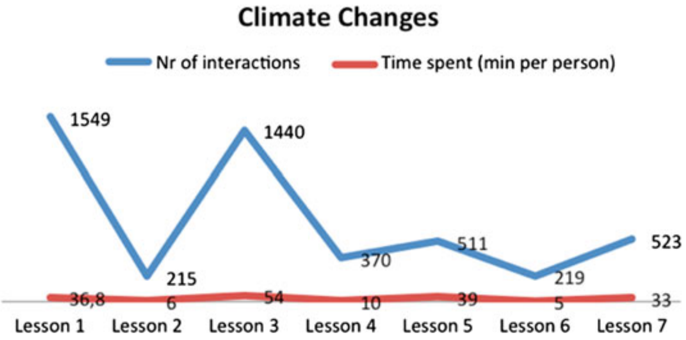


Fig. 8 Interactions and time spent on lessons (computer-supported inquiry MOOC/climate change MOOC)

extra within course settings; also, it may hint that course facilitators initiated some interactions every second week; and it may also suggest that some lessons were less interesting for participants.

5 Discussion

The analysis in the MOOCs shows some pedagogical neutrality of the platform—based on data that were analyzed, it can be said that EMMA platform supports different pedagogical designs of MOOC. MOOC could be xMOOC, which focuses on content consuming and could be also cMOOC where participants actively communicate and construct new knowledge together.

LA dashboards need to evaluated to find out in which way it supports the learning experience in EMMA platform and also how MOOC facilitators plan their pedagogical interventions of MOOCs to next iterations of the course: How do they make sense of the data, and how it is integrated to course design process.

LA data could be combined with the data of the participants' surveys in the platform: What expectations do they have when they enter to the MOOC and what is their learning path during the course.

6 Future Work

The development of the learning analytics framework and practical implementation also depends on the developments in the field, shared experiences from different projects and initiatives using similar approaches. The issues of data standardization, data collection, and analysis that go beyond the platform, specification, and use of particular verbs are the issues that also influence the EMMA platform LA developments. There is also a strong need for theoretical frameworks and connecting learning analytics data to specific learning scenarios and pedagogies, evaluation of LA dashboards based on specific questions, and understanding the *what* and *how* of the learning processes.

References

Buckingham Shum, S. (2012). Learning analytics. UNESCO policy brief enhancing teaching and learning through educational data mining and learninganalytics: An issue brief U.S. Department of Education Office of Educational Technology.

Chatti, M. A., Lukarov, V., Thüs, H., Muslim, A., Yousef, A. M. F., Wahid, U., & Schroeder, U. (2014). Learning analytics: challenges and future research directions.

Clow, D. (2013, April). MOOCs and the funnel of participation. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 185–189). ACM.

Daniel, J. (2012). Making sense of MOOCs: Musings in a maze of myth, paradox and possibility. *Journal of interactive Media in education*, *2012*(3).

Drachsler, H., & Kalz, M. (2016). The MOOC and learning analytics innovation cycle (MOLAC): a reflective summary of ongoing research and its challenges. *Journal of Computer Assisted learning, 32*(3), 281–290.

Engeström, Y. (2001). Expansive learning at work: Toward an activity theoretical reconceptualization. *Journal of Education and Work, 14*(1), 133–156. doi:10.1080/13639080020028747.

Jonassen, D. H., & Ronrer-Murphy, L. (1999). Activity theory as a framework for designing constructivist learning environments. *Educational Technology Research and Development, 47*(1), 61–79.

Kevan, J. M., & Ryan, P. R. (2016) *Experience api: flexible, decentralized and activity-centric data collection. Technology, knowledge and learning* (pp. 1–7).

Powell, S., & Yuan, L. (2013b). MOOCs and open education: Implications for higher education.

Reich, J. (2015). Rebooting MOOC research. *Science, 347*(6217), 34–35.

Tweet-Tales: Moods of Socio-Economic Crisis?

Grazia Biorci, Antonella Emina, Michelangelo Puliga, Lisa Sella and Gianna Vivaldo

Abstract The widespread adoption of highly interactive social media such as Twitter, Facebook, and other platforms allows users to communicate moods and opinions to their social network. Those platforms represent an unprecedented source of information about human habits and socio-economic interactions. Several new studies have started to exploit the potential of these big data as fingerprints of economic and social interactions. The present analysis aims at exploring the informative power of indicators derived from social media activity, with the aim to trace some preliminary guidelines to investigate the eventual correspondence between social media indices and available labour market indicators at a territorial level. The study is based on a large data set of about four million Italian-language tweets collected from October 2014 to December 2015, filtered by a set of specific keywords related to the labour market. With techniques from machine learning and user's geolocalization, we were able to subset the tweets on specific topics in all Italian provinces. The corpus of tweets is then analysed with linguistic tools and hierarchical clustering analysis. A comparison with traditional economic indicators suggests a strong need for further cleaning procedures, which are then developed in detail. As data from social networks are easy to obtain, this represents a very first attempt to evaluate their informative power in the Italian context, which is of potentially high importance in economic and social research.

G. Biorci
CNR-Ircres (Genova), Genoa, Italy
e-mail: grazia.biorci@ircres.cnr.it

A. Emina · L. Sella (✉)
CNR-Ircres (Moncalieri), Moncalieri, Italy
e-mail: lisa.sella@ircres.cnr.it

A. Emina
e-mail: antonella.emina@ircres.cnr.it

M. Puliga · G. Vivaldo
IMT (Lucca), Lucca, Italy
e-mail: michelangelo.puliga@imtlucca.it

G. Vivaldo
e-mail: gianna.vivaldo@imtlucca.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_19

Keywords Big data • Social media • Twitter • Hierarchical clustering • Unemployment

1 Introduction and Motivation

In last years, the enhancing development in big data science provided tools for collecting and analysing an unprecedented amount of information about human habits, socio-economic interactions, and collective decision-making. Data from search engines, search query volumes, Internet users, and so on provide useful information for predicting collective behaviour (Goel et al. 2010; Choi and Varian 2012; Moat et al. 2014).

Bentley et al. (2014) specifically explore the use of massive data from highly interactive social network platforms that allow users an instantaneous and pervasive communication of moods and opinions to their social networks (e.g. Twitter and Facebook), as a new information source to investigate collective decision-making. In a different perspective, many other studies conceive social media data as genuine fingerprints of socio-economic interactions, such as local economic development and the relation between mobility fluxes and unemployment (Eagle et al. 2010; Llorente et al. 2015).

Focusing on labour market flows, different big data sources and approaches are currently under exploration. Internet job search query indices have been extensively exploited to enhance the nowcasting of contemporaneous economic activity in Israel (Suchoy 2009), Italy (D'Amuri 2009), Germany (Askitas and Zimmermann 2009), and the USA (Choi and Varian 2009). Antenucci et al. (2014) propose a social media index of job loss for the USA, derived from counts of job-related expressions in Twitter data. Their real-time index fits an interesting tracking of initial claims for unemployment insurance data, which performs better predictions than both consensus forecast and lagged data, thus showing that tweets incorporate pieces of information which are not reflected elsewhere.

Arising from their experience, this paper presents a preliminary approach to analyse Italian-language Twitter data, aiming at investigating critical issues in the Italian labour market. In particular, an integrated approach is described that blends data science, textual/linguistic, and statistical techniques with the aim to trace some preliminary guidelines to investigate eventual correspondence between social media indices and available labour market indicators at a provincial level.

Analyses are performed on a large corpus of about three million Italian-language tweets, dated from October 2014 to December 2015 and filtered by a set of specific keywords that are semantically related to the labour market. By means of machine learning and users' geolocalization techniques, the tweets are subset in all Italian provinces. Given both the wide semantic richness of Italian language and the not truly satisfactory performance of automatic feature selection algorithms (Antenucci et al. 2013), the corpus is then analysed by domain knowledge linguistic techniques, with the double aim of inspecting the corpus general contents and moods and of extracting textual signals that potentially correlate with socio-economic indicators of job lack. Since preliminary statistical cluster analysis on that signal fits unclear comparison with traditional economic indicators, several noise cleaning procedures are developed, including users' detection strategies.

The rest of the paper is organized as follows: after a brief corpus description, a linguistic and textual knowledge domain analysis is presented that explores job- and unemployment-related issues. Then, count variables from unemployment-related tweets and unemployment rates at a provincial level are compared by a hierarchical clustering approach. Finally, the effect of noise is assessed, and advanced corpus cleaning procedures are proposed as a starting point for further analysis.

2 Corpus Description

The corpus is a collection of tweets from Twitter recorded from October 2014 to December 2015 using the Twitter Stream API (https://dev.twitter.com/streaming/public) filtered using a bag of words (BoW now on) that consists of Italian words such as contracts, (un)employment, job, layoff, young, govern, wage, act, workers union, and the names of some important political leaders. The BoW considers also several topics of interest to the Italian labour market, such as *articolo* 18 and jobsact, that are, respectively, an entry on the Worker's Statute about lay-off, and a job market reform proposed by Mr. Renzi, the present Italian Prime Minister. Each term of the BoW is extracted by case insensitive procedures, and hashtag versions are considered too (i.e. renzi, #Renzi).

The initial collection contains more than 12 millions tweets with texts in several languages that are filtered in two ways: (a) looking to the "language" field given by Twitter (a machine learning system able to distinguish among several languages in real time as all tweets are marked with the recognized language), (b) using a tool for language detection based on a large multilingual machine learning training data set (langid https://github.com/saffsd/langid.py). This double check allows to reduce the risk to misinterpret the language: assuming that Twitter has a precision of 90–95% and our tool langid has a comparable precision (90% as declared in the guide), this means that with the double check we are now misinterpreting 1 in 100 tweets.

Moreover, we were interested in selecting the geographical features of the tweets focusing on the location, that is, declared by the user at the time of his/her Twitter subscription. This feature must not be confused with the actual user's geolocation, that is, his/her actual position when the smartphone GPS system is on. In fact, for our purposes the GPS position can be strongly misleading, since it captures mostly the travel situations rather than the user's provenance.

The location field is declared by the users in the 20–30% cases, and it can be parsed to extract a large fraction of reliable locations. For instance, the location Lucca refers to a city in Italy, and it is clear and unequivocal, so we accept it; the location Italy or Tuscany is instead refused, being too generic. The geolocation process is made using a *gazetteer*, i.e. a long list of geographical Italian cities and villages, their coordinates, and their administrative subdivisions. Since our work focuses on the Italian provinces (i.e. subdivisions of regions), each tweet geolocation is associated with the correspondent province. The procedure is refined for common mistakes such as capital letters (i.e. ROME instead of Rome) and punctuation removal. A conservative approach is preferred, i.e. only the most safe users' locations are kept. This choice reduces the corpus to a final dimension of 3.209.715 tweets, including repeated tweets (retweets and reposts). The tweets were generated by a total 250.743 unique users.

3 Knowledge Domain: Linguistic and Textual Analysis

The corpus consists, thus, of a portion of lexicon which is heavily connoted to the domain of *work* in all its possible declinations, so *job/work* are our superordinated words. We started our queries testing expressions concerning *job loss* and *unemployment* domains.

For this purpose, we

(a) set a new BoW,[1] a selected series of linguistic patterns, words and syntagms, which are more likely expected to describe events concerning job loss, firing, and unemployment;
(b) verified their presence in the selected tweets;
(c) analysed semantic value and salience of the co-occurrences in the tweets;
(d) drafted a sort of twitter-thesaurus centred on *disoccupazione* (unemployment) and *lavoro* (job/work).

As first survey on social mood concerning unemployment and loss of job, we chose to start our test with the stemmed words *disocc** and *lavor**, following the hypothesis that the semantic values of the co-occurrences around such node words might suggest adequate hints. Within these subcorpora, we read and listed all the co-occurrences attested in the co-texts. We proceeded in grouping words (nouns, verbs, and adjectives) according to their semantic domain and observing their frequency and salience among the subcorpora. We were then able to shape a sort of twitter-thesaurus centred on mood expressions concerning unemployment and job, so we could highlight the main topics of tweets clustered under conceptual superordinated words (Tognini–Bonelli 2001).

By inquiring the terms in the co-texts and by figuring, quite significantly, their semantic domain, it has been possible to define a sort of descriptive grid of this portion of the corpus. In particular we:

[1] *perso il lavoro*, *perdere il lavoro*, *perdere il posto di lavoro*, *perderlo*, *senza lavoro*, *mancanza d* lavoro*, *mancanza lavoro*, *manca lavoro*, *manca il lavoro*, *lavoratori a casa*, *lasc* a casa*, *rest* a casa*, *giovani a casa*, *licenzi**, *disocc**.

- produced a list of concordances of the node words identified as salient and semantically important for our aims. The length of the concordance string was decided to be 140 characters, which is exactly the length of a standard tweet. We had, so, a complete, and significant, visualization of the co-texts;
- obtained a list of collocations, in which words are displayed, which are more frequent or statistically more probably close to the node word "disocc*". In this way, syntagms and co-occurrences of "disocc*" emerge and connote the salience of each node word;
- tested both collocations and clusters of node words and compared the results with those obtained by a close reading of the concordances.

From our text analysis and linguistic point of view, the problems we had to face may be resumed in two main topics: the software used and the significant presence of noise. As regard as the chosen text analysis software, *AntConc*, although innovative and usually well performing, it showed some constraints in creating and exploiting sub-partitions of the corpus and in exporting data. The presence of noise, conversely, was due to different causes: the automatic/robot retweeting, the presence of the emoticons codes, appearing as a series of letters/characters producing non-words, the presence of many symbols and diacritical signs used for emphasis, and rhetoric purposes.

4 Statistical Clustering: Preliminary Comparisons with Economic Indicators

As a first step, the geolocalized tweets of the 110 Italian provinces were analysed by hierarchical clustering[2] methodologies, in order to extract a common behaviour in terms of their Twitter activity. At this preliminary stage, the weekly provincial Twitter activity was computed by counting the raw occurrences of the stemmed word *disocc** inside each unique Twitter per province per week.[3] As a first result, we observed the lack of convergence of clustering algorithms, even at a quite high iteration rate, suggesting a serious instability in detecting the suitable number of possible communities present in the data set. Just few clusters were detected. In particular, the most populated provinces Rome and Milan emerged as isolated and dominating components, pointing that weekly count data need to be normalized for the province Twitter population. Since these data are unavailable, the total amount of Twitter users for each province was estimated by the province population itself,

[2] Hierarchical clustering was performed by Ward (1963) *minimum variance*, *complete,* and *single linkage* methods (Murtagh and Legendre 2014). The distance matrix was computed by *Euclidean* metric. Clusters uncertainty was assessed at the 95% c.l., following the approach of Shimodaira (2004) and Suzuki and Shimodaira (2006).

[3] Raw counts were standardized to zero mean and unit variance in order to ensure their reciprocal comparability.

under the hypothesis that the users are equally distributed among Italian provinces, i.e. it exists a direct proportionality between the population of a given province and its Twitter users. As the number of unique users changes in time for seasonal effects, we took the monthly average number of unique users per province, and we plotted it against the 2015 province population, revealing the presence of a strong linear relationship. Thus, two further weakly variables were computed: the number of unique users per estimated Twitter users (*users/Twitter_users*), and the amount of total weekly counts of *disocc** per estimated Twitter users (*counts/Twitter_users*).

A first visual Cattell (1966) *Scree test* on both *users/Twitter_users* and *counts/Twitter_users* variables suggested that the number of clusters was not trivial to assess, since no evident structural break clearly emerged between major and minor (trivial) factors. Hierarchical clustering was then performed.

According to preliminary results, *counts/Twitter_users* records were not stable to changes in the clustering method. Single and complete algorithms were unable to isolate significant components, while *Ward's* methodology detected four clusters at the 95% confidence level (Fig. 1, central panel).

The analysis was slightly more robust for *users/Twitter_users* records. A small cluster including Avellino and Parma provinces turned out to be more stable to algorithm changes, while just *Ward's minimum variance* method was able to detect other three clusters at the 95% c.l. (Fig. 1, right panel). Some correspondences clearly emerge between the two normalized variables.

The visual mapping of unemployment rates at provincial level in December 2014 (Fig. 1, left panel) showed the well-known partition in low-, medium-, and high-unemployment areas that does not have any clear counterpart in the two proposed social media indices. In fact, the *counts/Twitter_users* clustering (Fig. 1, central panel) shows a sort of bipartition between low- (blue) and high-unemployment (red) areas, but some low-unemployment Northern regions fall in the red cluster. On the contrary,

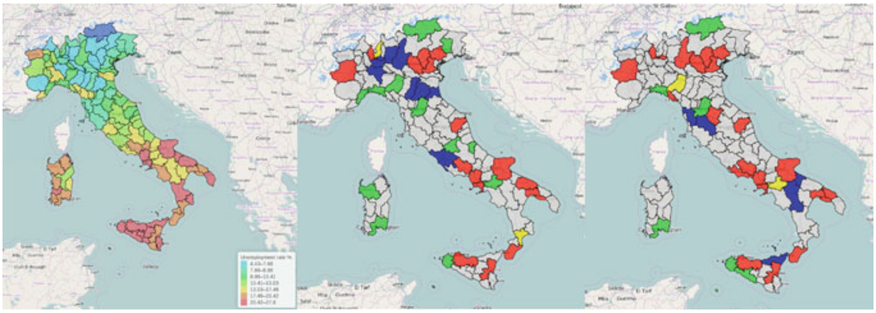


Fig. 1 Unemployment statistics versus Twitter unemployment indicators. *Left panel* official unemployment rates per province, Dec. 2014. *Source* our elaboration on Istat data. *Central panel* hierarchical clustering on counts/Twitter_users. *Right panel* hierarchical clustering on users/Twitter_users. *Source* our elaborations on Twitter data. Algorithm: Ward's (1963) minimum variance

the second indicator (*users/Twitter_users*, Fig. 1, right panel) does not show any sort of polarization in low- and high-employment areas.

5 Discussion and Preliminary Conclusions: Advanced Filtering Issues

This study focuses on statistical analyses of unemployment-related tweets in Italy, with the aim of exploring the informative power of indicators computed from social media activity in describing well-known socio-economic phenomena. Due to a large noise, preliminary clustering results were stable.

In fact, the initial cleaning procedure generating the unemployment-connoted subcorpus was scarcely able to describe geographical characteristics according to the well-known patterns of unemployment in Italy. Instead, data tended to be mixed, and hierarchical clustering results were not robust across algorithms. Hence, preliminary analyses suggested the need for further advanced filtering.

To improve the original filtering procedure, we investigated several tweet features, especially the interarrival time of two consecutive tweets for each user, that undercover the eventual presence of automatic retweeting systems (bots). The hypothesis is that no human user can tweet with a rate of one tweet per second or lower. As a matter of fact, the Twitter platform is particularly suited for automatic systems that can repost or retweet contents to increase the chances of spreading a message to a large audience. This characteristic can affect the descriptive/predictive power of signals extracted from Twitter data, when the analysis aims at measuring socio-economic phenomena from the echo they have in people conversations. In fact, in the corpus we identified job posting and press agencies tweets, as well as tweet reposts (retweets or reply) from automatic systems. Figure 2 (right panel)

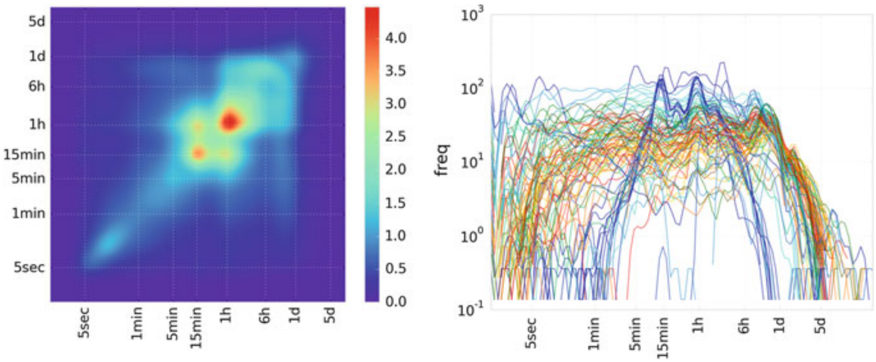


Fig. 2 Interarrival time plots. *Left panel* tweets interarrival time map across users. *Red-orange* regions represent in ln scale higher tweet activity for the corresponding interarrival time couple. *Right panel* tweet frequencies versus interarrival times for each user. Colours are proportional to individual average frequency. *Source* our elaboration on Twitter data

shows the distribution of interarrival times for each user. The curves display different behaviour: while most recurrent blue peaks refer to users that tweet on average every 15 min or 1 h, other users are tweeting every few seconds and in large numbers. These users are very likely to be bots, adding nonnegligible noise to the corpus.

Another class of users eventually generating noise is the professional users, such as press agencies, that usually issue posts at fixed times (e.g. 1 h). This behaviour is represented by the largest red area in the interarrival map plot (Fig. 2, left panel). We can make the data set cleaner removing this kind of users and their tweets, too.

Finally, as further filtering improvement, we could isolate and look at the peaks in the volume of tweets in time. When a new law or act is announced in the media, users react with a higher rate of comments. This phenomenon is known as *agenda setting* and can be used to evaluate the diffusion of the same news across different areas in Italy. We can measure the response to each tweet peak in terms of intensity (number of tweets) and persistence (decaying rate of each peak). A check of the geographical distribution of those parameters can be interpreted as a measure of interest of each topic and in each part of the country. The signal of the peaks is more clean than the baseline signal, that is, likely to be more noisy and less focused on the labour market topics. However, the peak signal is strongly dependent on the specific issue discussed by the media and this can be misleading in turn. We can easily identify the topic of the peak just looking at the media tweets that are responsible for the initial peak surge.

References

Antenucci, D., Cafarella, M. J., Levenstein, M., Ré, C., & Shapiro, M. D. (2013). Ringtail: Feature selection for easier nowcasting. *WebDB*, 49–54. Retrieved April 29 2016 from http://www-cs.stanford.edu/people/chrismre/papers/webdb_ringtail.pdf.

Antenucci, D., Cafarella, M., Levenstein, M., Ré, C., & Shapiro, M. D. (2014). Using social media to measure labor market flows. *NBER Working Paper Series,* w20010. National Bureau of Economic Research.

Askitas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *German Council for Social and Economic Data (RatSWD) Research Notes*, 41.

Bentley, R. A., O'Brien, M. J., & Brock, W. A. (2014). Mapping collective behavior in the big-data era. *Behavioral and Brain Sciences, 37*(1), 63–76.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 245–276.

Choi, H., & Varian, H. (2009). Predicting initial claims for unemployment benefits. Retrieved April 29, 2016 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.549.7927&rep=rep1&type=pdf.

Choi, H., & Varian, H. (2012). Predicting the present with google trends. *Economic Record, 88,* 2–9.

D'Amuri, F. (2009). Predicting unemployment in short samples with internet job search query data. *MPRA Paper*, 18403. University Library of Munich, Germany.

Eagle, N., Macy, M., & Claxton, R. (2010). Network diversity and economic development. *Science, 328*(5981), 1029–1031.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences, 107*(41), 17486–17490.

Llorente, A., Garcia-Herranz, M., Cebrian, M., & Moro, E. (2015). Social media fingerprints of unemployment. *PLoS ONE, 10*(5), e0128692.

Moat, H. S., Preis, T., Olivola, C. Y., Liu, C., & Chater, N. (2014). Using big data to predict collective behavior in the real world. *Behavioral and Brain Sciences, 37*(1), 92–93.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *Journal of Classification, 31*(3), 274–295.

Shimodaira, H. (2004). Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Annals of Statistics, 32,* 2616–2641.

Suchoy, T. (2009). *Query indices and a 2008 downturn: Israeli data*. Bank of Israel. Research Department.

Suzuki, R., & Shimodaira, H. (2006). Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics, 22*(12), 1540–1542.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work.* Amsterdam: John Benjamins Publishing.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58,* 236–244.

The Sentiment of the Infosphere: A Sentiment Analysis Approach for the Big Conversation on the Net

Antonio Ruoto, Vito Santarcangelo, Davide Liga, Giuseppe Oddo, Massimiliano Giacalone and Eugenio Iorio

Abstract In the Network Society the use of hashtags has become a daily routine for the participation on the Big Conversation Iorio and Ruoto (Nessun tempo, 2015). Designated by a 'hash' symbol (#), a hashtag is a keyword assigned to information that describes it and aides in searching. Hashtags are now central to organize information on Social Networks. Hashtags organize discussion around specific topics or events and they are becoming an integrated part of the Infosphere, the whole informational environment constituted by all informational entities. The sentiment analysis of Hashtags shared on the Big Conversation can return a possible snapshot about the sentiment shared by users. Scope of this work is to present an application of sentiment analysis on the Italian hashtags of mainly social networks as part of the 'Infosphere'. This analysis returns a semantic sentiment report about the hashtags shared by the users of the social networks, that can produce a semantic sentiment trend about users. This approach could be applied to every language simply changing the sentiment thesaurus used.

A. Ruoto · D. Liga
iInformatica S.r.l.s., Corso Italia 77, Trapani, Italy

A. Ruoto · E. Iorio
University of Naples 'Suor Orsola Benincasa', Naples, Italy

M. Giacalone
Department of Economics and Statistics, University of Naples 'Federico II', Naples, Italy

G. Oddo · V. Santarcangelo (✉)
Centro Studi S.r.l., Zona Industriale, Buccino, Italy
e-mail: santarcangelo@dmi.unict.it

V. Santarcangelo
Department of Mathematics and Computer Science, University of Catania, Catania, Italy

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_20

1 Introduction

This research paper aims to determine a methodology for investigating and mapping the emotional level of the Italian Big Conversation on the Net as a part of the larger Infosphere level. More specifically, this paper will provide an example of this methodology applied to the Instagram platform, and in particular to Italian-speaking users, with the following objectives:

- Assess what direction the generated emotional forces are going;
- Understand what are the main emotional projections being transmitted;
- Assess what are the ten most important emotional polarizations.

Sentiment analysis of text is a well-known technique in Natural Language Processing. The idea is that some words hold positive or negative meanings. For example the word 'good' might have a positive score '+2' and the word 'terrible' negative '−3'. Each word in a social network caption, added by its owner, is converted into a corresponding score, and all we have to do is to sum them up to get the overall mood. Sentiment analysis is based on sentiment thesaurus, that is customized for the target language. For the Italian language there are two thesaurus: AIN Thesaurus Santarcangelo et al. (2015) and SentiStrenght. The first is the most complete thesaurus for Italian language also for the number of words considered, for the difference between adjectives and intensifiers and also because it includes a semantic information for each word Santarcangelo et al. (2015). In the Philosophy of Information (PI) the term Infosphere denotes the entirety of the Information environment. Coined from the term 'biosphere', this neologism has been first introduced in the literature by Luciano Floridi, one of the most influent academics of the PI. According to Floridi, the infosphere is 'the semantic space composed by the entirety of documents, agents, and their operations', where 'documents' refers to any kind of data, information and knowledge that has been codified and achieved in any semiotic form; 'agents' refers to any system that is able to interact with an independent document (e.g. individuals, organisations, robot softwares); 'operations' refers to any action, interaction and transformation that an agent can carry out, and that at the same time can be regarded as a 'document'. All the above constitutes an environment in which organisms are able to develop, as if they were interconnected cells Floridi (2011). Considering todays Network Society, in which we all live, it is quite clear that the Big Conversation on the Net is currently an essential part of the infosphere. The widespread of the Internet, and the access to mobile devices, digital media, along with a wide range of social platform, fostered the development of interactive and horizontal networks of communication that are able to connect local and global spheres at any time. The communication system of the industrial era revolved around mass media, characterized by unidirectional one-to-many mass dissemination of information. Conversely, the basis of our Network Society revolves around a global system of horizontal communication capable of generating multimodal exchanges of information and many-to-many interactions (both synchronously and asynchronously). Following the consolidation of Manuel Castells mass self-communication paradigm, individuals have increasingly got used to such new means of communication and they have generated their own mass communication system, made of sms, blog, social networks and messages exchanged on all sorts of instant messaging platforms. It is a huge, multilinguistic and multicultural communication environment Castells (2009). In this sense, the Net has become a sort of place where everybody can say something; an infinite room in which a privatization of the public sphere occurs, an intimism that alters the conventional ideas of ‘public’ and ‘private’. If we consider the public sphere like a room in which public opinion develops, the analysis of the current Big Conversation on the Net (influenced by events and agenda-setting topics) can reveal interesting aspects of todays major transformations, providing an insightful prospective on the future mechanisms involved in the generation of common sentiment and collective/individual imagination. However, it should be considered that public opinion derived from the Big Conversation on the Net is giving way to a sort of emotional opinion. In particular, the effects of information overload and emotional sharing are going to turn the public debate into a debate that seems to be more emotional than sensible Lovink (2010), Carr (2010). Users do not pay attention to information unless it promotes their interest, enthusiasm, fear, anger, disgust. As a consequence, it seems that the only effective message is the emotional one, which originates from something strongly emotional. In this way, within the mass self-communication, the amplification of the emotional sphere becomes the foundation on which information can be spread; and this spread follows the rules of the emotional contagion. To a certain extent, it is the confirmation of what the philosopher David Hume argued almost three centuries ago: reason is the slave of passion, and not the reverse Westen (2008). Therefore, if we become aware of the fact that behind social behaviours there are processes that are able to alter social emotion, an action of emotional intelligence would help us understand the extent of such alteration, the way in which it influences common sentiment and the role it plays in the creation of collective and individual consciousness.

2 Methodology

In order to map the emotions circulating within the Big Conversation on the Net, this paper resorts to OSINT methodology, gathering information provided by publicly available sources. More specifically, considering a period of almost 6 years (from 10th October 2010, launch date of Instagram in Italy, to 10th January 2016), we have classified the most commonly used Italian adjectives that appear in the hashtags of Italian-speaking users of Instagram (with a frequency that is greater than or equal to 100.000) according to an ‘emotional scoring system’. Each hashtag has been assigned an ‘emotional score’ on the basis of the AIN Thesaurus Pilato et al. (2015), one of the most exhaustive Italian thesaurus for the Sentiment Analysis. This thesaurus assigns to each adjective an emotional polarity score which ranges from negative to positive, according to the following scoring system:

- + 2 (very positive)
- + 1.5
- + 1 (positive)
- + 0.5
- 0 (neutral)
- −0.5
- −1 (negative)
- −1.5
- −2 (very negative)

In addition to this, each hashtag has been classified according to the emotion classification system introduced by the American psychologist Robert Plutchik (2002) (Figs. 1, 2, 3 and 4).

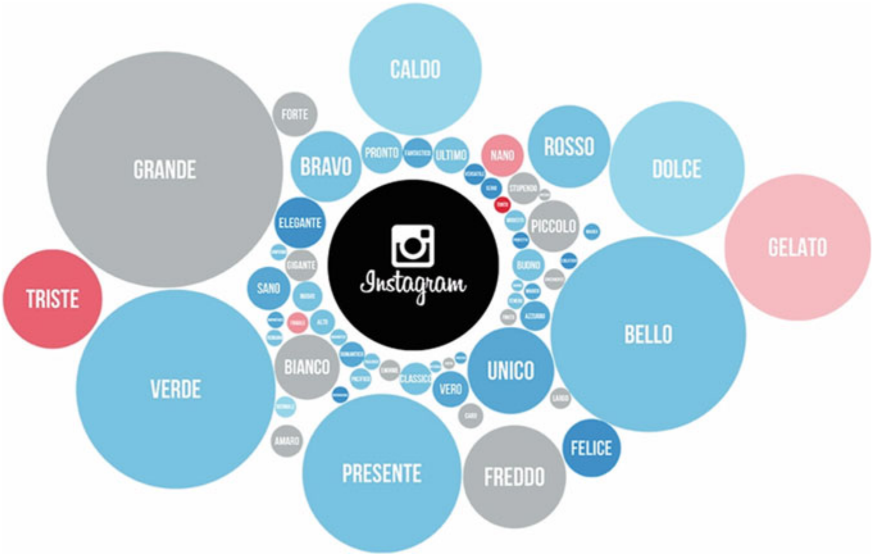


Fig. 1 Instagram sentiment hash cloud

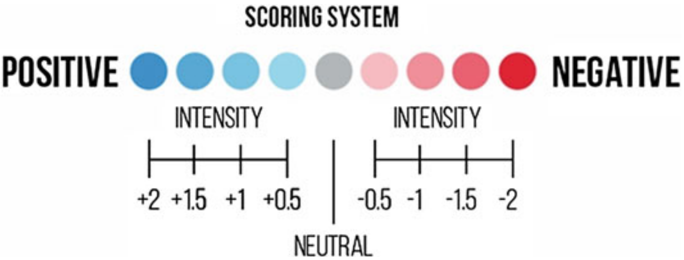


Fig. 2 AIN scoring system

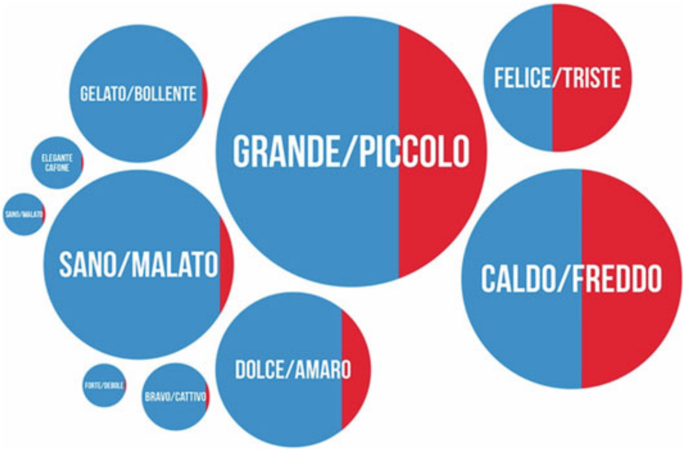


Fig. 3 Instagram top sentiment polarizations

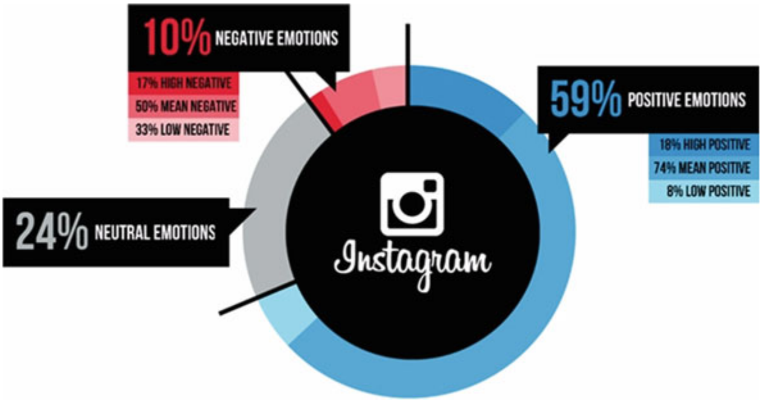


Fig. 4 Instagram results

3 Discussion on the Results

In short, the study suggests that Italian users of Instagram use the social network to:

- transmit emotional polarizations that are usually positive: of all the examined hashtags, those that have been used more frequently show more positive emotional polarity, whilst on average very few hashtags have a negative emotional polarity;
- generates a social environment in which the most represented emotional categories are those linked to the spheres of ‘anticipation’ and ‘joy’; conversely, emotions linked to the ideas of ‘anger’ and ‘disgust’ are almost non-existent.

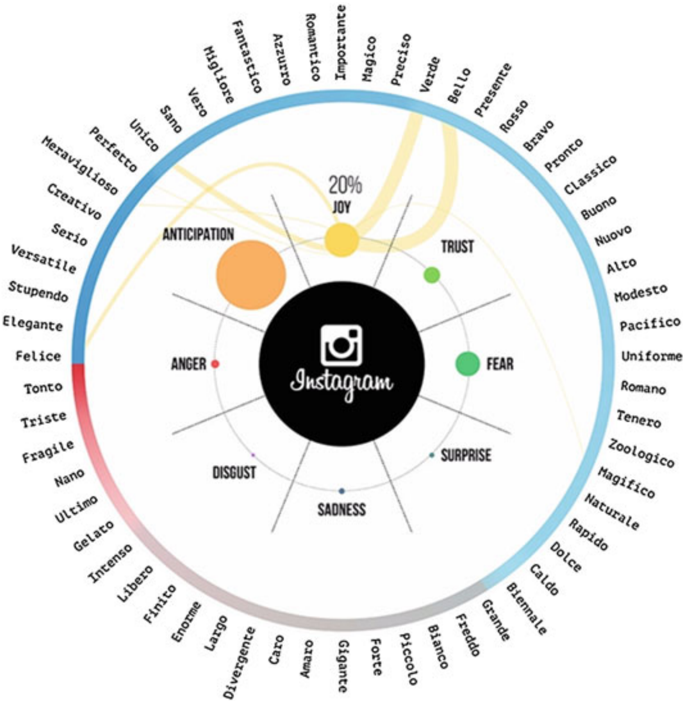


Fig. 5 Hashtag Plutchik analysis

These results should be interpreted taking into account the following key observation: the form in which contents are presented and the architecture of the social networking environment are more important than contents themselves. In other words, communication processes turn into storytelling. Due to the nature of the means, the narratives transmitted by Instagram users seem to serve one specific purpose: the digital packaging of the Self. In this sense, the users content creation leads, more or less consciously, to a sort of self-marketing in which users main objective (declared or not) is to paint himself in the best light possible in the market of social relationships. In a society in which you exist only if you appear, this general trend seems to respond to a need for people to ‘perform themselves’ (Figs. 5, 6).

4 Conclusions

It is not surprising that the emotions expressed in hashtags of our study tend to be positive. It is simply a seductive strategy, a sort of negotiation: by transmitting a positive emotion and narrative, users receive more positivities (e.g. likes). In this way, it seems that Instagram leaves little room for a truly critical debate on reality. Further

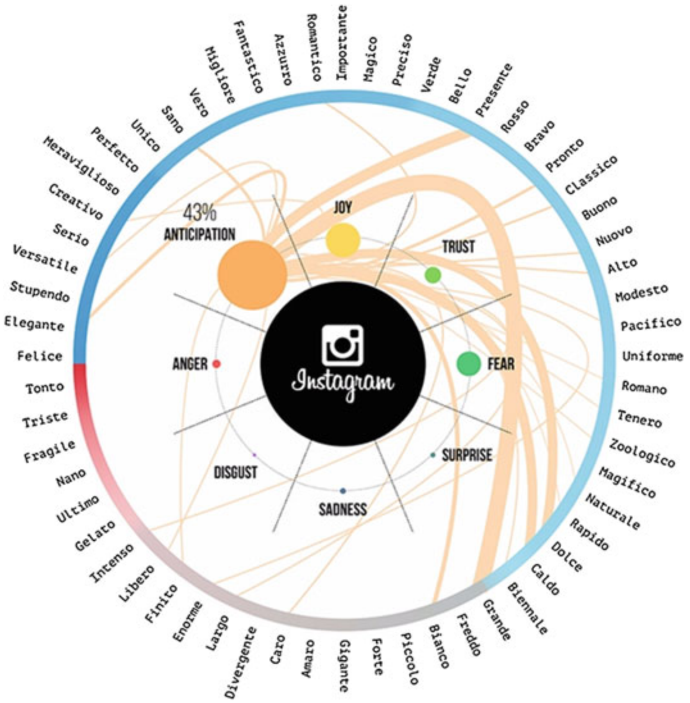


Fig. 6 Hashtag Plutchik analysis—anticipation

evidence of this is also provided by the predominance of a homophile reasoning in users' dynamics. In this regard, likeability ends up being not only the true yardstick when developing a narrative strategy aimed to take centre stage on social networks, but also a major driving force in the creation of common sentiment.

References

Carr, N. (2010). The shallows: What the Internet is doing to our brains. W. Norton & Company

Castells, M. (2009). *Communication power*. New York: Oxford University Press, Oxford.

Floridi, L. (2011). *Information a very short introduction*. Oxford: Oxford University Press.

Iorio, E., & Ruoto, A. (2015). *Nessun tempo*. Nessun Luogo: La comunicazione pubblica italiana all'epoca delle Reti.

Lovink, G. (2010). *Networks without a cause: A critique of social media*. Cambridge: Polity.

Pilato, M., Santarcangelo, G., Santarcangelo, V., & Oddo, G. (2015). AIN Thesaurus, RCE MULTIMEDIA

Plutchik, R. (2002). *Emotions and life: Perspectives from psychology, biology, and evolution*. Washington, DC: American Psychological Association.

Santarcangelo, V., Pilato, M., et al. (2015). *An opinion mining application on OSINT for the reputation analysis of public administrations*. Bari: Choice and preference analysis for quality improvement and seminar on experimentation.

Santarcangelo, V., Oddo, G., Pilato, M., Valenti, F., & Fornaro, C. (2015). Social opinion mining: an approach for Italian language. SNAMS2015 at FiCloud2015.

Westen, D. (2008). The political brain: The role of emotion in deciding the fate of the nation. PublicAffairs

The Promises of Sociological Degrees: A Lexical Correspondence Analysis of Masters Syllabi

D. Borrelli, R. Serpieri, D. Taglietti and D. Trezza

Abstract Our work consists of a secondary data analysis of Sociological (LM-88) Masters Syllabi displayed in MIUR dedicated website (http://www.universitaly.it). For each course, we regrouped the available data in two main dimensions: the first is inherent the structure of the formative paths of students and graduates; the second one is inherent the Syllabi "promises", regarding the prospected acquisition of cognitive and professional competencies. We choose to work on textual data in order to achieve two different goals. First of all, we can outline recurrent semantic areas through a lexicometric analysis. Then, we will be able to discuss the relations between the above-mentioned dimensions, pointing out the distance between the promises and the existing structure. We think that this approach will be helpful to light up what is often ignored in the evaluation of master degrees, namely the content of official syllabi. We appreciate this analytical perspective because it shows how the sociological courses supply is widely differentiated and evaluate the distance between this varied supply and the actual condition of the graduates.

Keywords Lexical correspondence analysis · Textual data · Sociology courses

1 Introduction

Since the end of the last century, in Italy, higher education institutions (HEIs) have been subjected to many structural reforms, concerning their governance, funding, evaluation and recruitment systems. In the political discourse, the value of accountability has been emphasized as the most important principle for legitimating these reforms, according to the widespread narrative of new public management rules (Moini 2015).

D. Borrelli
University of Salento, Lecce, Italy

R. Serpieri · D. Taglietti · D. Trezza (✉)
University of Naples "Federico II", Naples, Italy
e-mail: domenico.trezza@unina.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_21

During 2012, MIUR designed the dedicated website http://www.universitaly.it, in order to provide all users of HEIs with detailed information about the syllabi of all courses supplied by national universities. Since it goes online, all aspects of HEIs should be easily at the hand of their stakeholders. A particular effort has been devoted to make all degree courses the clearest and the most accessible as possible to their potential students, in terms of main promises about educational aims and opportunities of job placement. What is at stake is a better fitting between supply and demand of professional competencies, in order to enhance careers guidance of the graduates and to minimize the possible skill mismatching in the labour market.

This essay aims to shed light on the self-presentation strategies offered by the 21 Italian master degrees in sociology, comparing them with the real formative structure of each course (Facchini 2015). In particular, analysing the official syllabi displayed on MIUR website, for each course, we regrouped available data in two main dimensions: the first is inherent to the structure of the formative programme followed by students and graduates; while the second is inherent to syllabi promises, regarding the prospected acquisition of professional competencies and the potentialities in terms of placement.

We choose to focus mainly on textual data in order to achieve two different goals. First of all, through a lexicometric analysis, we are able to outline recurrent semantic areas among educational provisions of all 21 Italian master degrees in sociology. Then, we inquired the relation between above-mentioned dimensions, pointing out the distance between the promises and the structure of their educational curricula.

Even if the self-presentations of the master degrees are generally profile specific and are autonomous features for different learning processes and career options, we found that the design of courses in fact is relatively homogeneous, anyway more similar than the self-presentations would make us think. Our impression is that the differentiation in the self-presentations is more suited to the aims of academic marketing than to the needs actually related to educational and scientific processes.

Our essay consists of three sections: in the first, a short overview on the adopted methods is presented; in the second one, we will analyze the structure of programmes followed by students, in order to depict the concrete articulation of each course supplied by universities; and, finally, in the third part, we will show the application of textual analysis techniques to our research.

2 Methodological Framework: The Textual Analysis

In the last years, in order to meet the ever-growing need to analyze big textual data, it became necessary to develop new methods involving the use of semiautomatic techniques processing texts.

Quantitative textual analysis allows us both to carry out lexicometric operations for a first textual documents exploration, either to represent the forms on a factorial plane, by multidimensional statistic techniques. The aim is to detect latent semantic dimensions behind the text.

The aim of this work, as well as theoretical findings, is to show a set of methodological steps by LCA reaching three findings (Amaturo 1989):

- A summary of the information contained in the data;
- Display of multiple associations between words;
- Connection between text data and contextual data.

2.1 A Brief History of the Content Analysis

The first example of content analysis was the Sion songs analysis being carried out by Swedish Church—in the XVII century—who believed that they were heretics. However, starting from Laswell studies in the 1930s about political symbols, quantitative and systematic documents' research has been gaining more and more relevance (Amaturo and Punziano 2013). *Classification unit* becomes key elements in the textual research. According to Rositi (1988), there are three *classification units*: one that is recognizable within the text as it corresponds with the signifier, i.e. a word. Another classification unit is not directly related to grammar and linguistic level, but it has an evidence quite high in the text. Finally, the third c.u. coincides with context unit, as if the document or the text were people to be interviewed.

Regarding this work, we have dealt with the first-type unit since examining the information content we have made a words lexicometric analysis. Moreover, Rositi distinguishes even the classification units of the first type. He identified four classification units—the word, the key symbol, the sentence and the theme. The simplest unit is the word. However, there is no agreement about the meaning of word in the content analysis.

According to some linguistic scholars, word is what defines a concept, while for others it identifies a graphic form, a simple character's sequence delimited by spaces and points. The symbol key can not only be the single term, but also a words composition having a specific meaning. The sentence corresponds with a part of the text. Finally, the theme refers an assertion about a text item that is operationalized by researcher to be quantitatively detected.

2.2 The Lexical Correspondence Analysis

This brief overview about quantitative research in the text analysis is useful to understand methodological context of our research.

On the basis of our research aims, we were interested in understanding what were the lexical strategies used by courses to introduce themselves, and especially how they relate with contextual data about the training offer. The Europeanization of the university system has led towards a standardization of textual contents ever stronger. This has certainly facilitated our content analysis work. As previously mentioned, the use of LCA technique was invaluable since it has allowed us not only an exploratory study on the textual content, but mainly it allowed us to connection between the textual data and contextual data—in our case the training offer ECS.

LCA is a multidimensional analysis technique applied to a text, and it was born from the correspondence analysis of the data analysis French school (Lebart and Salem 1988). Therefore, the originality of this kind of analysis is to apply a text to the CA technique.

The LCA data are organized in an array in which the rows are lexical forms of all documents and the columns are the same documents. Indeed, the single forms are categories of a variable “lexicon” while the documents are categories of a variable “textual corpus” (Tipaldo 2007).

Through parameters such as Chi-square, the aim is to detect the similarity and dissimilarity between rows and columns.

With the aid of appropriate statistical techniques such as factor analysis, and thanks to the use of specific analysis software—In our case, the SPAD_T software—it was possible to graphically represent the associations between the rows and columns of the table on a plane defined by two factorial axes.

In the following sections, we will show the application of this technique, which revealed some interesting insights, especially in relation to the comparison of textual and contextual data.

3 Formative Structure of Courses

In this paragraph, we consider the first dimension in which we regrouped available data. Our aim is to analyze the structure of programmes followed by students, in order to depict the concrete articulation of each course supplied by considered universities.

To do so, we focus our attention on two kinds of secondary data inherent to the structure of programmes: the number of characterizing ECS for each disciplinary area (sociological, mathematical, anthropological, political and philosophical) and the number of ECS for each SSD in the sociological area. Through a descriptive analysis, we are able to calculate the frequency distribution of each disciplinary area and of each SSD on the total of characterizing ECS. In Fig. 1, we can observe the pie chart that shows the distribution of ECS among different disciplinary areas: we can see that 54% ECS are in the sociological area, while other four areas have quite equivalent weight (about 11–12%). In Fig. 2, then, we can observe the pie chart that focuses on the distribution of ECS internal to the sociological area. Here, we can see that 3 prevalent SSD (sps/07, sps/08 and sps/09) achieve the 74% of the total of sociological ECS, while other 3 SSD (sps/10, sps/11 and sps/12) share the remaining 26%.

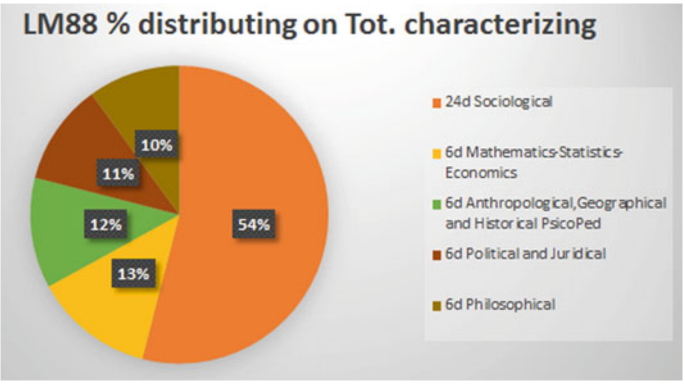


Fig. 1 Distribution of ECS among different disciplinary areas

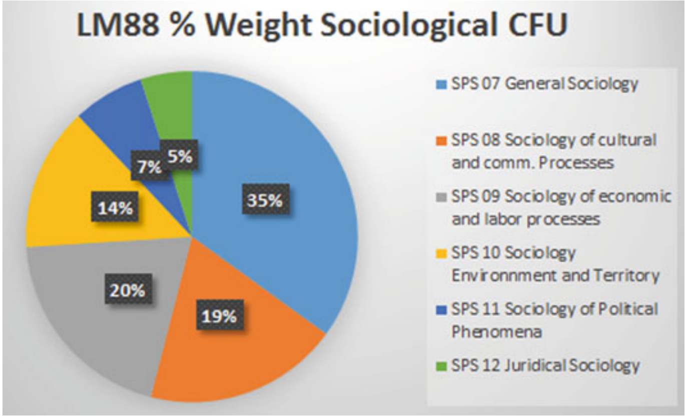


Fig. 2 Distribution of ECS in the sociological area

We try to operationalize the significance given by each course to each disciplinary area through such a categorization (Fig. 3). For the sociological area, considering the lower limit of compulsory ECS stated by a national law, we calculate exceeding ECS and categorize them in three intervals of 33.3 percentile width each. So, in the first interval, we find 7 low-sociological-content courses; in the second one, we find 8 middle-sociological-content courses; and, finally, in the third one, we find 6 high-sociological-content courses. For other areas, the lower limit of compulsory ECS stated by law reveals itself less pregnant, because exceeding ECS are homogeneous. So, we used a dichotomous categorization, distinguishing between courses with presence or absence of exceeding ECS in each disciplinary area and considering this like a proxy of the significance granted to it. In Fig. 4, we can see that 4 courses give more significance to the philosophical area, 6 to the political, 12 to the mathematical and 6 to the anthropological one.

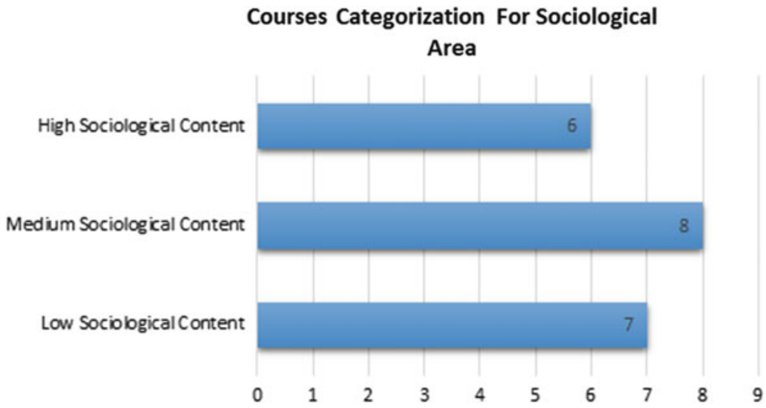


Fig. 3 Courses categorization for sociological area

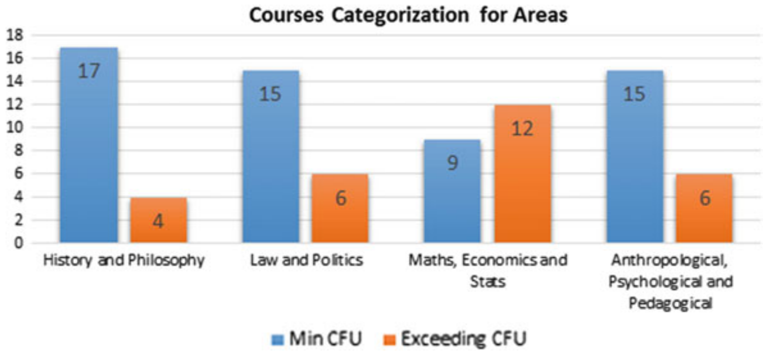


Fig. 4 Courses categorization for other areas

Finally, we narrow our attention on the sociological area and categorize courses in function of an "index of incidence" of each sociological SSD on the total of characterizing sociological ECS. This index, calculated per row, permits us to reveal the weight of each sociological SSD compared with others. To do so, we established four intervals by considering the whole distribution: by this way, we are able to weigh the relative significance of the SSD for each course. As shown in Fig. 5: the first interval, with value of zero, gathers those courses which give no significance to the specific sps; the second interval, with value between 0.1 and 19.9, gathers those courses which give little significance to the specific sps; the third interval, with value between 20 and 50, gathers those courses which give medium significance to the specific sps; and, finally, the fourth interval, with value more than 50, gathers those courses which give high significance to the specific sps. All these intervals represent contextual data that will be compared with textual data, analyzed in the next paragraph.

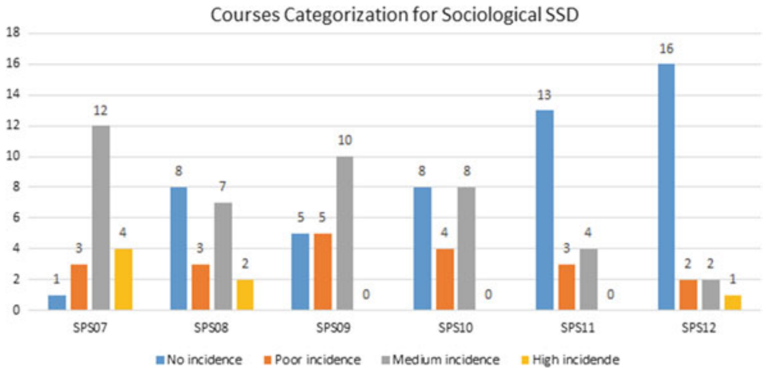


Fig. 5 Courses categorization for sociological SSD

4 The Courses Promises. a Data Sheets Content Analysis

The key step of our work is a content analysis of courses data sheets available on MIUR website. As previously mentioned, we analyze these data following three different statistical–methodological stages: the lexicometric analysis of textual forms, the lexical correspondence analysis and the cluster analysis.

This permits us to obtain a clear view about lexical strategies used by courses designers and to identify recurrent semantic profiles.

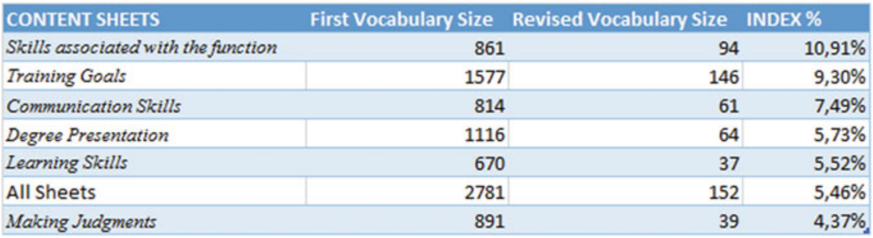
4.1 The First Exploration of the Textual Space—A lexicometric Analysis

In order to preliminary explore the textual material and to prepare it for the next analytical stage, we need a statistical analysis of all words in the *corpus*.

The first fundamental step is the textual pretreatment phase, after the vocabulary constitution, i.e. a list of all graphic forms—defined by delimiter scripts—with their occurrences.

The raw vocabulary had an overall amplitude of 2.781 several graphic forms, from which derived the first empirical evidence: the lexical amplitude of the *Training Goals* sheet—1.577 forms—is opposed to the *Learning Skills* sheet with 670 graphic forms. So, we need to do more preliminary actions to reduce the *corpus*.

The normalization allows us to delete duplicates due to accents and capital letters. The elimination of *empty forms* (Bolasco 1999)—i.e. all elements poor of sense, such as conjunctions, articles and prepositions—reduces the vocabulary amplitude, keeping only the interesting words.



CONTENT SHEETS	First Vocabulary Size	Revised Vocabulary Size	INDEX %
Skills associated with the function	861	94	10,91%
Training Goals	1577	146	9,30%
Communication Skills	814	61	7,49%
Degree Presentation	1116	64	5,73%
Learning Skills	670	37	5,52%
All Sheets	2781	152	5,46%
Making Judgments	891	39	4,37%

Fig. 6 Size vocabulary and “richness lexical” index

Giving to the lemmatization process, which allows to reduce each word to a unique grammatical form, and the lower limit setting of occurrences for each textual form—i.e. 4—we get the definitive vocabulary, with 152 several words.

It is hard to establish how much diversified the lexical structure of the sociology courses presentations is, due to the lack of substantial comparative data; however, the analysis of the single forms, by calculating the ratio of the number of words in the final vocabulary on the number of the initial one, might suggest a sort of lexical repetitiveness, with low percentages going from the 10.7% of the *skills associated with the function* section, to another Dublin descriptor, the *making judgments* (4.4%) (Fig. 6).

The most frequent words analysis of the several sections seems to confirm both the small vocabulary used and a substantial homogeneity in the content data sheets. For example, considering the first five more frequent words for each of the six sections, we can see how the term *social* is present in the top five of 4 sections and it is also the most recurrent form.

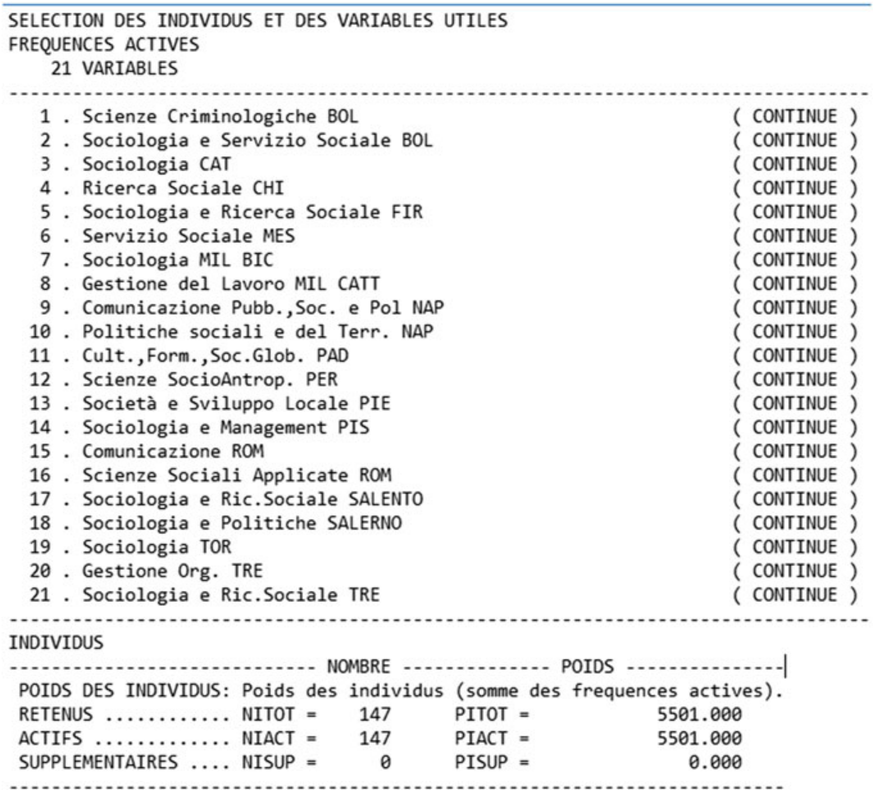
Words such as *competences* (in 5/6 sheets top five), *research* (4/6), *abilities*, and *knowledge* (3/6) suggest that the course information goes beyond the specificity of the single section.

4.2 A Textual Universe: Four Lexical Constellations

The quantitative text analysis gives us some empirical evidences that we will now study in deep by lexical correspondence analysis. Our aim is to understand which are the semantical structures underlying the LM88 course description.

This was possible through the graphical projection of extract factors, which allows us to visually analyze the textual points cloud and reveal the semantical associations. Also, if this is a technique that allows multidimensional analysis, we put in active—i.e. with contribution to the factors construction—just 2 variables: one corresponding to the lexicon, and so having words as modality, the other referring to the courses (Fig. 7).

Due to methodological reasons, it is opportune to put two points in illustrative, corresponding to the public, social and political communication and criminological



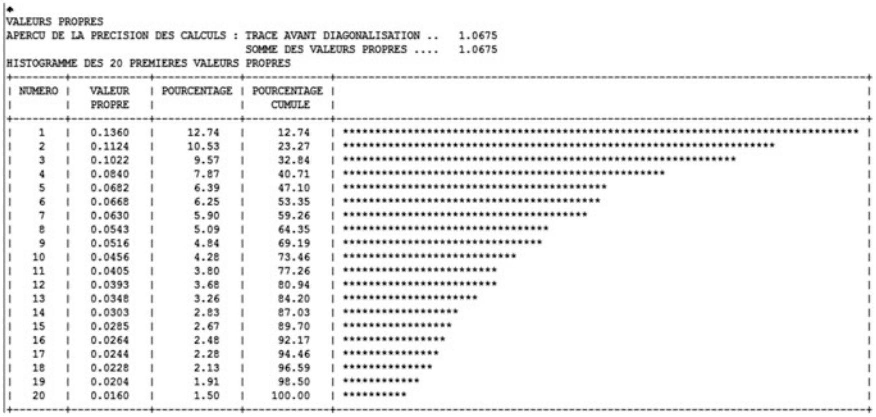
```
SELECTION DES INDIVIDUS ET DES VARIABLES UTILES
FREQUENCES ACTIVES
    21 VARIABLES
----------------------------------------------------------------------------------
   1 . Scienze Criminologiche BOL                                    ( CONTINUE )
   2 . Sociologia e Servizio Sociale BOL                             ( CONTINUE )
   3 . Sociologia CAT                                                ( CONTINUE )
   4 . Ricerca Sociale CHI                                           ( CONTINUE )
   5 . Sociologia e Ricerca Sociale FIR                              ( CONTINUE )
   6 . Servizio Sociale MES                                          ( CONTINUE )
   7 . Sociologia MIL BIC                                            ( CONTINUE )
   8 . Gestione del Lavoro MIL CATT                                  ( CONTINUE )
   9 . Comunicazione Pubb.,Soc. e Pol NAP                            ( CONTINUE )
  10 . Politiche sociali e del Terr. NAP                             ( CONTINUE )
  11 . Cult.,Form.,Soc.Glob. PAD                                     ( CONTINUE )
  12 . Scienze SocioAntrop. PER                                      ( CONTINUE )
  13 . Società e Sviluppo Locale PIE                                 ( CONTINUE )
  14 . Sociologia e Management PIS                                   ( CONTINUE )
  15 . Comunicazione ROM                                             ( CONTINUE )
  16 . Scienze Sociali Applicate ROM                                 ( CONTINUE )
  17 . Sociologia e Ric.Sociale SALENTO                              ( CONTINUE )
  18 . Sociologia e Politiche SALERNO                                ( CONTINUE )
  19 . Sociologia TOR                                                ( CONTINUE )
  20 . Gestione Org. TRE                                             ( CONTINUE )
  21 . Sociologia e Ric.Sociale TRE                                  ( CONTINUE )
----------------------------------------------------------------------------------
INDIVIDUS
------------------------------ NOMBRE -------------- POIDS ---------------|
 POIDS DES INDIVIDUS: Poids des individus (somme des frequences actives).
 RETENUS ............ NITOT =    147      PITOT =            5501.000
 ACTIFS ............. NIACT =    147      PIACT =            5501.000
 SUPPLEMENTAIRES .... NISUP =      0      PISUP =               0.000
--------------------------------------------------------------------------
```

Fig. 7 “Courses” variable modes

sciences courses. This is because the strong contribution to the two factors, due to the use of strongly characteristic forms, does not allow a good read of the graphic. The histogram of the factors in Fig. 8 suggests to consider the first three factors: according to scree test method we will consider only the first two of them, because even if they reproduce a low share of inertia (22%), they seem to support our reflections.

Before reading the graphic, which alone can lead to serious analytical distortions, a view of the absolute and relative contributions of the points-words leads us to think that the first factor (12% of Inertia tot. Reproduced) is relative to content that refer to the learning experience. Indeed, forms such as *Policies* (5.4), *Apprenticeshi*p (3.7), *Analysis* (3.3) and *Preparation* (3.1) make us think of two types of learning experiences: one oriented to the field analytical, insisting on theoretical formative activities as the analysis, design and evaluation.

The other most voted to the presentation of practical activities, which refer to the formation of the “role” through apprenticeships, internships and audits. The second axis (10.3%) in our opinion “intercepts” those contents that refer to the



```
◆
VALEURS PROPRES
APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION ..   1.0675
                                     SOMME DES VALEURS PROPRES ....   1.0675
HISTOGRAMME DES 20 PREMIERES VALEURS PROPRES
+--------+------------+------------+------------+----------------------------------------------------------------------------------+
| NUMERO |   VALEUR   |POURCENTAGE |POURCENTAGE |                                                                                  |
|        |   PROPRE   |            |   CUMULE   |                                                                                  |
+--------+------------+------------+------------+----------------------------------------------------------------------------------+
|    1   |   0.1360   |    12.74   |    12.74   | ******************************************************************************** |
|    2   |   0.1124   |    10.53   |    23.27   | ******************************************************************               |
|    3   |   0.1022   |     9.57   |    32.84   | ************************************************************                     |
|    4   |   0.0840   |     7.87   |    40.71   | *************************************************                                |
|    5   |   0.0682   |     6.39   |    47.10   | ****************************************                                         |
|    6   |   0.0668   |     6.25   |    53.35   | ***************************************                                          |
|    7   |   0.0630   |     5.90   |    59.26   | *************************************                                            |
|    8   |   0.0543   |     5.09   |    64.35   | ********************************                                                 |
|    9   |   0.0516   |     4.84   |    69.19   | ******************************                                                   |
|   10   |   0.0456   |     4.28   |    73.46   | ***************************                                                      |
|   11   |   0.0405   |     3.80   |    77.26   | ************************                                                         |
|   12   |   0.0393   |     3.68   |    80.94   | ***********************                                                          |
|   13   |   0.0348   |     3.26   |    84.20   | *********************                                                            |
|   14   |   0.0303   |     2.83   |    87.03   | ******************                                                               |
|   15   |   0.0285   |     2.67   |    89.70   | *****************                                                                |
|   16   |   0.0264   |     2.48   |    92.17   | ****************                                                                 |
|   17   |   0.0244   |     2.28   |    94.46   | **************                                                                   |
|   18   |   0.0228   |     2.13   |    96.59   | *************                                                                    |
|   19   |   0.0204   |     1.91   |    98.50   | ************                                                                     |
|   20   |   0.0160   |     1.50   |   100.00   | *********                                                                        |
+--------+------------+------------+------------+----------------------------------------------------------------------------------+
```

Fig. 8 Factors histogram

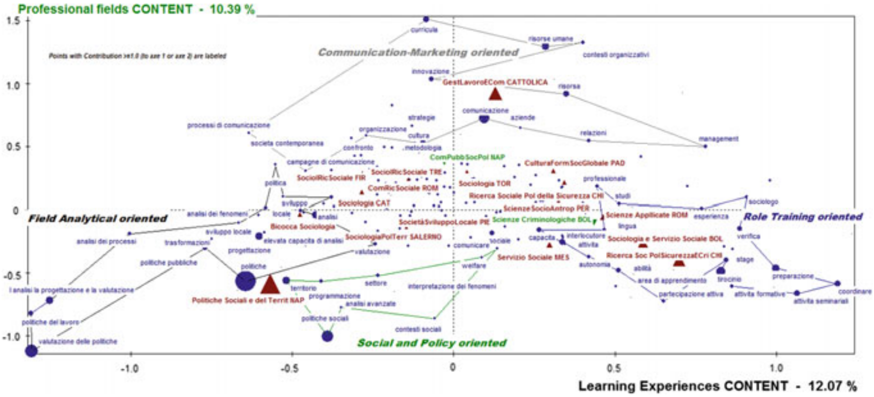


Fig. 9 Factorial plane—*lexical constellations*

"professionalizing" nature of the courses: on one side, there is a "constellation" semantically related to marketing issues and communication with terms such as *Communication* (6.4), *Human Resources* (4.0), *Organization* (3.1) and *Innovation* (2.8). It is in fact the area of courses in communication. The other semi-axis seems characterized by contents oriented to the representation of professional fields of social and territory policies—6.6 *Social Policy*, *Evaluation Policy* 3.6, *Territory and Context* 2.3. Fig. 9 clearly shows four profiles that appear as several "constellations". A different thing is the area that gravitates around the centre of the axes.

The factorial plane shows that the cloud of points of the courses will focus for the most part around the origin. This seems to follow the low lexical diversification in the course presentations. Observing the textual forms near the average profile, we realize that they—if we exclude the word "social"—are part of a little specific semantic universe for sociological field, but common for the university educational offer in general (*Work*, *Skills*, *Research*).

In the next section, we will see if the four lexical constellations described are confirmed by the clustering process.

4.3 The Cluster Analysis to Identify the Courses Lexical Profiles

The cluster analysis is the last statistical–methodological step of our analysis. The dendogram's cut (Fig. 10), obtained by hierarchical classification, leads to obtain 4 classes made up through the Ward method, that by acting on the inertia it aims to minimize decomposition of new groups. Such cut seems appropriate since it strikes a good balance in terms of the internal inertia for each group (28—21—35—16%). Through an analysis of the elements of four groups, we find that clustering partially confirms the evidence that emerged thanks to the LCA (Fig. 11):

- ***Knowledge Applied***: it represents the group with two-class masters whose semantic strategies aim to communicate their knowledge applied in the working world [*activities—interventions—participation—autonomy—research* are the words most characteristic of the group]
- ***Analysts***: it refers to the group that is oriented, at least semantically, to the classical methods to do research, which deal with data survey and data analysis [*competence—training—data—processing—research*]
- ***Communication-Marketing***: where the semantic strategy is linked to the marketing, business and communication themes [*learning—English_language—communication—organization—model*]
- ***Research-Evaluation***: is the group where the semantic content reflects the strong involvement in research activities aimed at the policy evaluation [*advanced skills—territory—policy—planning—evaluation*].

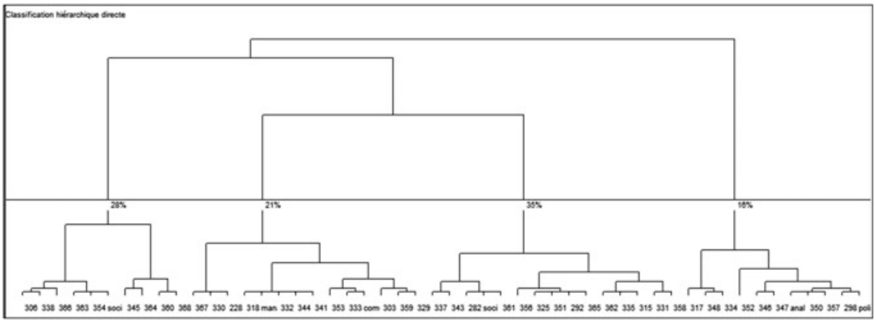


Fig. 10 Dendogram's cut

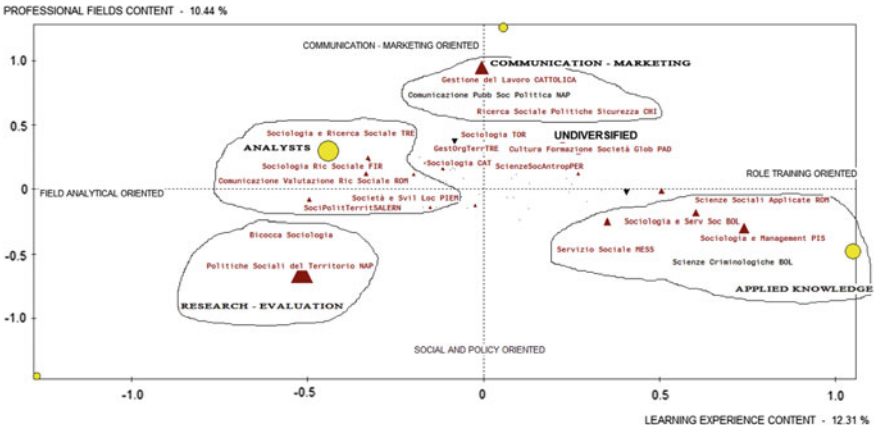


Fig. 11 Factorial plane—cluster and courses

Finally, in order to incorporate all the remaining elements near the origin of the axes, we decide, arbitrarily, to define a new group—called ***Undifferentiated***—that does not have its own lexical specificity.

So, these profiles represent an attempt of semantic structures synthesis that are behind the course presentations. We can say that are promises. So, how much are they different from the reality?

5 Conclusions

As we have pointed out, using the multiple correspondence analysis allowed us to detect five different types of lexical profiles among all Italian master degrees in sociology, as well as to compare them with some contextual data concerning the real features of their courses.

As emerged from our research, each one of five master degree's profiles would seem to promise a kind of educational provision quite well differentiated and targeted in terms of self-representations. If you compare these self-representation profiles with the design of their teaching programs, you can see that promises of master degrees are not actually confirmed as a whole.

Indeed, we obtained a new factorial plane (Fig. 12) among three variables, with two active variables—ECS sociological and ECS subject area—and a new variable, narrative profile type. It shows us approximately 38% of explained inertia. The main evidence comes from the first axis (21.5% of inertia) that seems to intercept the ECS incidence degree. However, the comparison with lexical profiles allows us interesting remarks.

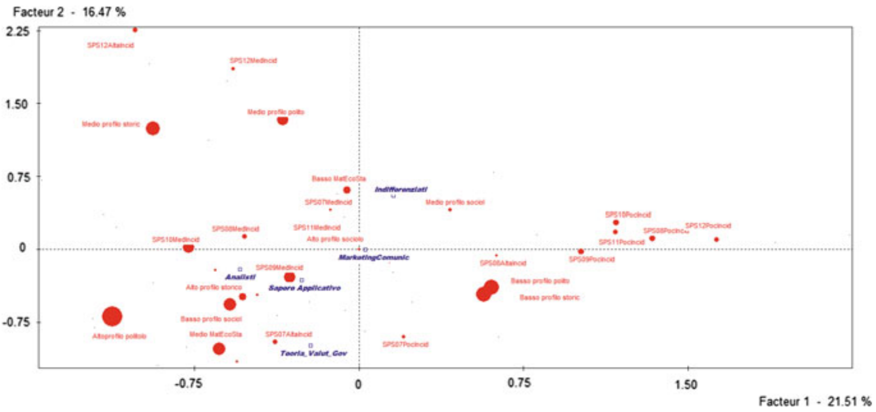


Fig. 12 Factorial plane—lexical profiles and ECS

As the graph points out, the difference among narrative profiles of the master degrees does not match an equally significant difference among the formative structure of the respective courses in terms of academic credits (generally speaking and, in particular, as regards the specific sociological courses).

We would like to emphasize three main findings of our study. First of all, in spite of the differences "promised" according to the self-representations, three of the five groups of master degrees (i.e. Knowledge Applied, Analysts and Research-Evaluation) present a very similar educational provision: in the lower right quadrant of the axis, we can see that all three groups are mostly characterized by many academic credits in history and in political science, as well as by a teaching sociological profile relatively low in terms of offered credits.

As part of the specific sociological credits, those of general sociology and economic sociology prevail for each of the three groups of master degrees.

Secondly, the master degrees whose courses lexical profiles outline a strong bent for marketing and communication studies, actually offer a formative structure quite generic, anyway not particularly oriented towards these specific disciplines, judging above all by the number of the related academic credits.

In almost every respect, their educational provisions occupy a middle placement among all 21 master degrees (we see them just in the middle of the graph): this means that, even though they represent themselves as outliers compared to all other courses, they de facto feedback the image of the most typical Italian master degree in sociology.

Finally, we would like to highlight that the group of master degrees we called undifferentiated for not showing up a course lexical profile so well characterized and defined, paradoxically are just those courses that differ most effectively from the others as regards their educational provision.

In conclusion, our research revealed that the formative design of different master degrees is anything but differentiated with respect to each other, anyway not enough to justify the different placements that their self-representations would imply on website “Universitaly”. This evidence would confirm that narrative emphasis on differences among the Italian master degrees in sociology are more suited to the aims of academic marketing than to well-founded motives of educational design.

References

Amaturo, E. (1989). *Analyse des données e analisi dei dati nelle scienze sociali*. Torino: Centro Scientifico Editore.

Amaturo, E., & Punziano, G. (2013). *Content Analysis: tra comunicazione e politica*. Milano: Ledizioni.

Bolasco, S. (1999). *Analisi multidimensionale dei dati*. Roma: Carocci.

Facchini, C. (2015). *Fare i sociologi. Una professione plurale tra ricerca e operatività.* Bologna: Il Mulino.

Lebart, L., & Salem, A. (1988). *Statistique textuelle*. Paris: Dunod.

Moini, G. (2015). *Neoliberismi e azione pubblica*. Il caso italiano. Roma: Ediesse.

Rositi, F. (1988). L’analisi del contenuto. In M. Livolsi & F. Rositi (Eds.), *a cura di La ricerca sull’industria culturale*. Roma: La Nuova Italia Scientifica.

Tipaldo, G. (2007). *L’analisi del contenuto nella ricerca sociale*. Edizioni Libreria Stampatori: Spunti per una riflessione multidisciplinare. Torino.

Part IV
Off-Line Data Applications

Exploring Barriers in the Sustainable Microgeneration: Preliminary Insights Thought the PLS-PM Approach

Ivano Scotti and Dario Minervini

Abstract Several sociological works have highlighted the non-technical barriers in the implementation process of green facilities and energy efficiency technologies. These researches have identified some of the main variables involved in the so-called sustainable transition process. They have also suggested some possible strategies to overcome the limits to promoting sustainable energy. Despite these scientific considerations, the dissemination of a microdistributed green generation —which seems to represent a deep shift in the current energy regime—is less developed compared to the renewable big facilities. Using statistics data from different sources on Italy, the aim of this work is to understand which dimensions appear to be important in overcoming non-technical barriers in the development of microdistributed green energy generation, in particular domestic photovoltaic plants and building insulation systems. Through the partial least squares path modelling (PLS-PM), we present some preliminary insight in exploring the non-technical barriers involved in the microgreen.

Keywords Green energy · Microgeneration · Non-technical barriers · PLS-PM

1 Introduction

Sustainable energy transition is a relevant issue in the contemporary sociological debate and scholars have developed analytical frameworks to scrutinize it. The multilevel perspective (Schot and Geels 2008) and the social practice theory (Shove 2012), for example, propose two different approaches to unfold the dynamic of the energy transition. Despite the differences, they both claim a pivotal relevance to the situated grass-roots experiences where transition takes place (as niche levels or everyday practices). Green innovations appear featured by specifics aspects

I. Scotti (✉) · D. Minervini
Department of Social Sciences, University of Naples Federico II,
Vico Monte Di Pietà 1, 80138 Naples, Italy
e-mail: ivano.scotti@unina.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_22

mutually intertwined with more standardized factors (i.e. regulations), it suggests that transition can be enacted within a very different socio-technical configurations. In the case of the co-provision model, for example, local actors can apply new exchange rules (Sauter and Watson 2007) obtaining more socio-environmental advantages compared with the spread of big green facilities (Seyfang et al. 2013).

Starting from these considerations, our aim is to reflect on the barriers to the development of the sustainable distributed generation in Italy through a very provisional study using the partial least squares path modelling (PLS-PM) approach. Some official statistics data related to the private microgeneration (small photovoltaic and solar thermal plants dedicated for self-consumption as an example of the distributed generation) will use to obtain preliminary insights about the role of social dimensions in the green transition. In this case, PLS-PM method appears an effective tool allowing to detect interaction among social dimensions involved in the transition process.

2 Barriers to the Microgeneration

What are the barriers to the development of a co-production distributed energy system in the case of private self-consumption? In the literature on the energy transition, we can identify two main research streams on this topic.

The first is focused on the policy, financial and technical dimensions in the microgeneration development. Scholars highlighted three factors that seem to obstacle the microgeneration diffusion: the poor financial aids provided to subsidize the technology costs, the administrative difficulties in obtaining these grants and a lack of information and accountability about output of microgeneration (Allen et al. 2008). Moreover, researches have shown different approaches within the population respect the microgeneration and the attitude to invest in it. In particular, young people seem more involved in the environmental issue, but they have to face their purchase power limits (Balcombe et al. 2013). If feed-in-tariff system consistently promotes and supports the transition, at the same time the reliability and transparency of the information about the tariffs is a crucial point in the construction of an institutional trust about energy policies (Simpson and Clifton 2015).

A second set of studies concerns the community and individual level. It focuses on the attitudes, social trust as well as communication aspects to explain the rate of microenergy technologies diffusion. Here, uncertainty on investments and technologies of microgeneration seems to be a significant barrier. Households have to choose between different options and have to be aware about the selective financial support dedicated to specific devices and solutions (Baskaran et al. 2013). Also, the environmental awareness and personal attitudes seem to play an important rule (Claudy et al. 2010). Young people, for example, because of their education and digital confidence, appear as a segment of population that can spread knowledge and information on microgeneration. Finally, the institutional weakness in promoting microgeneration seems to play a negative rule to development bottom-up experiences in the distributed energy production (Magnani and Osti 2015).

In short, looking to the literature, three main obstacles to the diffusion of the microgeneration can be pinpointed: (1) the financial constrains which reduce the propensity to invest in green technologies, (2) the personal attitudes, believes and competences on microgeneration and energy-environmental problems, (3) the presence/absence of promoting policies (i.e. feed-in tariff system or net metering for subsidizing microgeneration). The above-mentioned issues are connected in socio-technical practices and contexts, and, in what follows, a model is proposed as a plausible relational pattern assuming that social trust has a relevant impact in the adoption of microgeneration.

3 The Partial Least Squares Approach

The partial least squares (PLS) method is an approach to the structural equation modelling (SEM). Here, we report only the main arguments on the PLS, and for further insight, readers can refer to the scientific literature (Tenenhaus et al. 2005). PLS allows us to study cause–effect relationship models among latent variables (LV)—the not directly measured concepts, such as loyalty and satisfaction—operationalize by observable elements called manifest variables (MV). The PLS assumes that MVs contain information related to the LV blocks and through them is possible to obtain an approximate representation of the analytical categories and their relationships. The way to measuring LVs can be different: considering latent variables constructed by indicators (formative way) or supposing that indicators are caused by latent variables (reflective way). In the first case, we have a covariance-based method designed to confirm a theoretical model, in the second one is estimated the amount of variance explained to investigate the existence of relationship patters. Generally, the PLS approach uses a reflective strategy to identify LVs. From the SEM standpoint, PLS does not impose any distributional assumptions on the data. In this case, the proposed models are only considered approximation patterns with useful predictiveness. To do so, the PLS algorithm computes some parameters for the outer and inner model of SEM.

The inner model refers to relationship between the LVs and its MVs. PLS calculates weights to the latent variables, which indicate the MVs contribution to define the LVs. The algorithm also computes loadings as correlations between a latent variable and its indicators; if the block is homogeneous and unidimensional, the loadings are positive. Negative loadings indicate that manifest variable is not caused by latent variables and researchers can decide to exclude some MVs, dividing the not unidimensional block in sub-blocks or changing the measuring LVs (from reflective to formative). The solution is strictly related to the theoretical assumptions in any specific study. Three indicators are used to check block's homogeneity and unidimensionality. The Cronbach's alpha (which measures the internal consistency) and Dillon-Goldstein's rho (for block reliability) confirm the homogeneity if both are larger than 0.7, while a block can be considered unidimensional if, in the principal component analysis of a block, the first eigenvalue of its correlation matrix is higher than 1 and the others are smaller.

In the case of the inner model, PSL algorithm calculates path coefficients according to the model and related p-values, which measures the probability of "null hypothesis" for the relations. Through path coefficients detects the direct, indirect and total effect of each LVs in the cause-effect relationships, while the p-value shows if relationships between LVs are reliable or just result of chance.

The goodness and reliability of the relationship model in PLS method are generally estimated through four indexes. In particular, the commutability index estimates the goodness of the measured model, while the redundancy one shows the goodness of the inner model. For each structural relation, the R-squared reports the predictive ability of the model and the GoF—the goodness of fix (Henseler and Sarstedt 2013)—indicates the global model reliability.

4 Data and Relationship Model

As above reported, the paper aims is to explore the relationship among relevant dimensions detected in the scientific literature on the energy transitions, which affect the spread of the green microgeneration. PLS method seems worthy to help in this effort and we have chosen the *plspm* package of R software (Sanchez 2013) to perform the analysis. Some data from official statistics were taken into account to the analysis, however, these refer to the twenty Italian regions and this little sample can allow only a provisional and preliminary study. Reasonably, the selected variables among those available appear as good approximation to the dimensions emerged in the literature. In particular, we consider five dimensions (or LVs):

1. Social relationship: four MVs, rate of person who provided free aid to people (FRE), rate of people for civic and political participation (CIV), percentage of people who participate in associations (PAR), association funding (FUN). As it is known, these variables are linked to the social capital concept, which appear to promote economic development, institutional high performance and environmental engagement (Dasgupta and Serageldin 2000). Here, we assume that high social trust levels are indirectly connected with the adoption of microgeneration.
2. Environmental concern, three manifest variables: people concerned about the extinction of animals and plants (ANI), about natural disasters caused by human activities (DIS), the destruction of forests (FOR) and persons concerned on the natural resources depletion (RES). As Palm and Tengvard (2011) stress, people with high ecological concern are more liable to adopt microgeneration.
3. Economic well-being, 4 manifest variables: annual average household income (HIN), GDP per capita (GDP), labour force participation rate (LAB) and percentage of "satisfied" and "very satisfied" people for their economic situation

(SOD). Following Scarpa and Willins (2010), an implicit assumption could be that in regions with high economic well-being people are willing to invest in microgreen energy generation.

The ISTAT (Italian Institute of Statistics) data were used for these first three blocks.[1] Other two latent variables are composed by the ENEA and GSE data, as well as, Legambiente information.[2] Specifically:

4. Local energy policies, 3 MVs from ENEA: adoption of mandatory energy regulations (REG), spread of the incentive policies of the green energy (INC) and municipal regulations to promote green energy (MUN). As many works have shown (i.e. Genus 2012), appropriate incentive policies strongly push the adoption of green energy.[3]
5. Green microgeneration, 3 manifest variables from the GSE and Legambiente: small photovoltaic plants ($\leq$ 3 kW) per 1,000 inhabitants (PVS), solar thermal panels per 1,000 inhabitants (HEA) and rate of photovoltaic MW rooftop per 1,000 inhabitants (ROO). The solar technologies have been chosen because they seem to represent the microgeneration phenomenon in an appropriate way.

The relationship model we want to explore (Fig. 1) suppose that the "social trust" is the exogenous latent variable because, according to the patter, is not influenced by other LVs.

On the contrary, the "environmental concern", "local energy policies", "economic well-being" and "green microgeneration" are endogenous latent variables because in the proposed model they are caused by the exogenous ones. In short, the model of the causal relationship we want to obtain information according to our hypothesis and assumptions is reported in the Fig. 1.

5 The Analysis Results

In the PLS method, we need to take in consideration firstly the outer model. The blocks homogeneity and unidimensionality condition have to be checked. In our case, all LV blocks have good indexes despite the "local energy policies" one. In that case, the Cronbach's alpha and Dillon-Goldstein's rho are lower than 0.7. Considering its loadings, only one is negative and it refers to the manifest variable MUN (municipal regulations to promote green energy). This aspect seems to suggest that not mandatory legislative initiatives promoted by local authorities are

[1] Further information on data and metadata, visit the ISTAT web page: http://www.istat.it/en/.

[2] ENEA is an Italian Government-sponsored research and development agency, which undertakes studies on energy and other research areas. GSE is the state-owned company that promotes and supports renewable energy sources and the energy efficiency in Italy. Legambiente is an important Italian environmentalist association, which supports reports on green energy in Italy.

[3] On these data we refer to the ENEA report: *Energy Efficiency Annual Report 2013*.

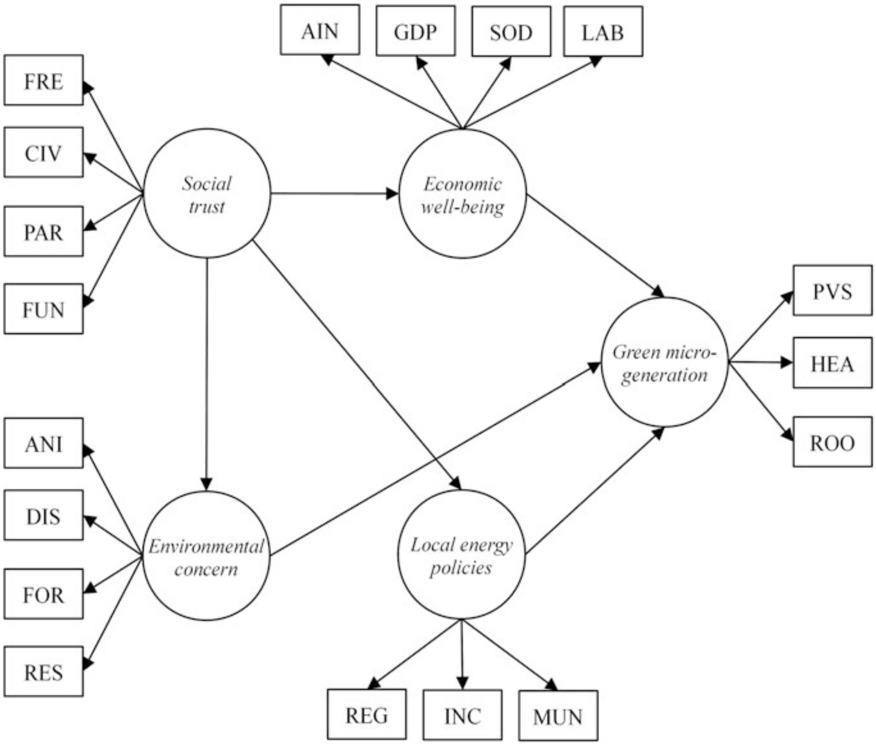


Fig. 1 PLS-PM model

Table 1 Homogeneity and unidimensionality of blocks

	C's alpha	D-G's rho	1st eigenvalue	2nd eigenvalue
Social trust	0.96	0.97	3.59	0.23
Environmental concern	0.84	0.90	2.74	0.73
Economic well-being	0.97	0.98	3.68	0.19
Local energy policy	0.70	0.87	1.54	0.46
Green microgeneration	0.91	0.94	2.54	0.33

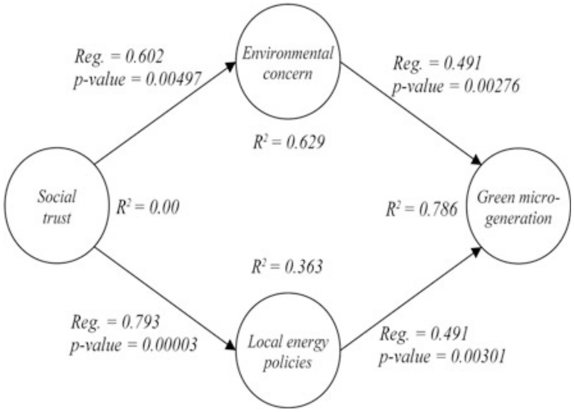
less relevant to spread microgeneration. Using few cases (Italian regions) seems not reasonable to perform a comparing group or a REBUS—Response-Based Unit Segmentation—analysis (Esposito Vinzi et al. 2008) to better explore the role of MUN. In a next research, sub-regional data would allow us to do a deeper analysis taking in account the microlevel and the rule of local regulations on the sustainable energy transition paths. For that, the MUN is deleted and the blocks homogeneity and unidimensionality condition is met (Table 1).

Loadings and communalities also need to be observed. Loadings report the correlation between latent variables and its indicators, while communalities measure the part of the variance between them. Despite we have good indexes (all loadings higher than 0.8 and all communalities higher then 0.7), to better specify the model, we have decided to deleted the DIS manifest variable in the "ecological concern" block because its communalities is low. It does not change the theoretical assumption of the model, but in this way, outer pattern appears statistically better defined.

About the inner model, PLS shows the path coefficients and related p-values. In particular, the algorithm output reports not significant p-values related to the structural relation between "economic well-being" and the "green microgeneration" LVs. The outcome seems to suggest two reasonable explanations on that first, green energy policies can reduce the importance of the financial resources to invest in microgreen technologies; second, the cost of green energy equipments (in our case, small photovoltaic and solar thermal) have decreased significantly in last years and the financial aspects has become less relevant. Considering these aspects, it seems useful to delete the "economic well-being" LV and perform again the PLS algorithm. In this second case, the main indexes of the inner model (Fig. 2) are quite good despite the R-squared of the "Local energy policy" is low.

First of all, the indexes of the "Green microgeneration" block, the final step of the path, are significant. Its R-squared that refers to the model predictiveness is 0.786; the redundancy, which calculate the goodness of the pattern, is 0.667 and the block communality, estimating the goodness of the measured model, is 0.848. Also the goodness of fix index is good (0.704), confirming the reliability of the model. In short, our provisional relationship model (despite we used a database composed by official statistics, which contain not completely coherent information for our purposes) seems to allow us to confirm the relevance of the social trust in the process of green transition. On the one hand, it seems to support a higher ecological awareness/concern that encourages the adoption of microgeneration; on the other, it favours the implementation/application of green energy policy at local level, which reduces the relevance of the financial aspects on the microgeneration adoption.

Fig. 2 Inner model of the PLS pattern



6 Conclusions and Further Investigations

In this paper, we have argued about the barriers that preventing the diffusion of the sustainable distributed generation in Italy. A PLS-PM approach was adopted in a preliminary attempt to investigate relations between different analytical dimensions of the phenomenon. So far, a very small sample (corresponding to the twenty Italian regions) and secondary sources (based on official national statistics), were used in the study, and results seem to confirm the high relevance of the social dimension in the green microgeneration transitions. Our results appear consistent with findings from the social sciences literature on the microgeneration. In particular, the social trust, which influences the environmental awareness and the adoption of supporting policies in local contexts, appears as pivotal dimensions on the diffusion of distributed generation options. In this case, we have also tried to explore the interaction among some dimensions involved in the transition process, framing a simple, but reliable, model connecting those dimensions and their weight in the propose relationship pattern.

The results of the analysis suggest that energy policies on the microgeneration should be focused not only in increasing the financial propensity to invest in it, but also to promote social trust (namely, social capital), which encourages a grass-roots transition path. However, our preliminary research can offer only suggestions on this aspect and further investigation are definitively needed. In particular, two questions seem to be relevant in the explanation of the microgeneration transition and have to be investigated more in detail. First, which role can be played by those regulations on sustainable energy promoted by local authorities (i.e. Municipalities) and how they can foster the social capital? Second, can we notice territorial differences when microgeneration is not explained by a “shared” level of local social capital and why? Indeed other socio-technical dimensions (for instance, the urban configuration, technological lock-in, etc.) are deeply involved in the process, this is why a more sophisticated and complex relational model, based on new ad hoc variables and more detailed cases, is needed.

References

Allen, S. R., Hammond, G. P., & McManis, M. C. (2008). Prospects for and barriers to domestic micro-generation: A United Kingdom perspective. *Applied Energy, 85,* 528–544.

Balcombe, P., Rigby, D., & Azapagic, A. (2013). Motivations and barriers associated with adopting microgeneration energy technologies in the UK. *Renewable and Sustainable Energy Reviews, 22,* 655–666.

Baskaran, R., Managi, S., & Bendig, M. (2013). A public perspective on the adoption of microgeneration technologies in New Zealand: A multivariate probit approach. *Energy Policy, 58,* 177–188.

Claudy, M., Michelsen, C., O’Driscoli, A., & Mullen, M. R. (2010). Consumer awareness in the adoption of microgeneration technologies. An empirical investigation in the Republic of Ireland. *Renewable and Sustainable Energy Reviews, 14,* 2154–2160.

Dasgupta, P., & Serageldin, I. (Eds.). (2000). *Social capital: A multifaceted perspective*. Washington: World Bank.

Esposito Vinzi, V., Trinchera, L., Squillacciotti, S., & Tenenhaus, M. (2008). REBUS-PLS: A response-based procedure for detecting unit segments in PLS path modelling. *Applied Stochastic Models in Business and Industry, 24,* 439–458.

Genus, A. (2012). Changing the rules? Institutional innovation and the diffusion of microgeneration. *Technology Analysis & Strategic Management, 24*(7), 711–727.

Henseler, J., & Sarstedt, M. (2013). Goodness-of-fit indices for partial least squares path modelling. *Computational Statistics, 28*(2), 565–580.

Magnani, N., & Osti, G. (2015). Does civil society matter? Challenges and strategies of grassroots initiatives in Italy's energy transition. *Energy Research & Social Science, 13,* 148–157.

Palm, J., & Tengvard, M. (2011). Motives for and barriers to household adoption of small-scale production of electricity: Examples from Sweden. *Sustainability: Science Practice & Policy*, 7 (1), 6–15.

Sanchez, G. (2013). *PLS path modeling with R*. Berkeley: Trowchez Editions.

Sauter, R., & Watson, J. (2007). Strategies for the deployment of micro-generation: Implications for social acceptance. *Energy Policy, 35,* 2770–2779.

Scarpa, R., & Willins, K. (2010). Willingness-to-pay for renewable energy: Primary and discretionary choice of British households' for micro-generation technologies. *Energy Economics, 23*(1), 129–136.

Schot, J., & Geels, F. W. (2008). Strategic niche management and sustainable innovation journeys: Theory, findings, research agenda, and policy. *Technology Analysis & Strategic Management, 20*(5), 537–554.

Seyfang, G., Park, J. J., & Smith, A. (2013). A Thousand Flowers Blooming? An Examination of Community Energy in the UK. *Energy Policy, 61,* 977–989.

Shove, E. (2012). Energy transitions in practice: the case of global indoor climate change. In G. Verbong & D. Loorbach (Eds.), *Governing the energy transition: reality, illusion or necessity?* (pp. 51–74). London: Routledge.

Simpson, G., & Clifton, J. (2015). The Emperor and the cowboys: the role of government policy and industry in the adoption of domestic social microgeneration systems. *Energy Policy, 81,* 141–151.

Tenenhaus, M., Esposito Vinzi, V., Chatelin, Y.-M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis, 48*(1), 159–205.

Individual Disadvantage and Training Policies: The Construction of “Model-Based” Composite Indicators

Rosanna Cataldo, Maria Gabriella Grassia, Natale Carlo Lauro, Elena Ragazzi and Lisa Sella

Abstract In evaluating a policy, it is fundamental to represent its multiple dimensions and the targets it affects. Indeed, the impact of a policy generally involves a combination of socio-economic aspects that are difficult to represent. In this study, regional training policies are addressed, which are aimed at closing the huge gaps in employability and social inclusion of Italian trainees. Previous counterfactual estimates of the net impact of regional training policies reveal the need to observe and take into account the manifold aspects of trainees’ weaknesses. In fact, the target population consists of very disadvantaged individuals, who tend to experience difficult situations in the labour market. To overcome this shortfall, the present paper proposes Structural Equation Modelling (SEM) that considers the impact of trainees’ socio-economic conditions on the policy outcome itself. In particular, the *ex ante* human capital (HC) is estimated from the educational, social and individual backgrounds. Next, the labour and training policies augment the individual HC, affecting labour market outcomes jointly with individual job-search behaviour. All these phenomena are expressed by a wide set of manifest variables and synthesised by composite indicators calculated with Partial Least Squares SEM (SEM-PLS). The construction of the SEM is appraised and applied to the case of trainees in compulsory education.

R. Cataldo · M.G. Grassia · N.C. Lauro
Università Federico II (Napoli), Naples, Italy
e-mail: rosanna.cataldo2@unina.it

M.G. Grassia
e-mail: mgrassia@unina.it

N.C. Lauro
e-mail: clauro@unina.it

E. Ragazzi · L. Sella (✉)
CNR-Ircres, Moncalieri, Italy
e-mail: lisa.sella@ircres.cnr.it

E. Ragazzi
e-mail: elena.ragazzi@ircres.cnr.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_23

Keywords Impact evaluation · Labour policies · Composite indicators · Structural equation models

1 Introduction and Motivation

The work based on SEMs we present in this paper assesses the entry and exit conditions of individuals undertaking a training course. This statistical exercise has to be put in the perspective of our wider work, which is the evaluation service for the European Social Fund Regional Operating Programme (POR-FSE) 2007–2013. In this context, we are required to perform a yearly placement analysis, aimed at estimating the net impact in terms of employment of training policies, based on a quasi-experimental design (Benati et al. 2017).

The ESF promotes a wide range of social and labour policies aimed at enhancing social inclusion in the member states. Within this framework, we were asked to concentrate on a subset of training policies sharing some common characteristics. All courses considered in the evaluation exercise were long (ranging from some 300 h of modules up to 3 year courses), full time, provided a final certificate (either a professional qualification or a specialization), included a compulsory traineeship period and were principally addressed at the unemployed. For the sake of generality, courses for highly disadvantaged groups (e.g. prison inmates or disabled individuals) were not included in this study. Three main types of training were represented: qualification courses for young people, basic qualification courses addressed at migrants and low-education level adults, and advanced specialization courses addressed at highly educated adults with qualifications not appreciated by employers. On the whole, these are mainly disadvantaged students, who have experienced a difficult school education and/or labour market transition and have chosen vocational training in the hope of finding new job opportunities.

In this paper, we concentrate on training courses for young people eligible for compulsory education (CE, Ragazzi and Sella 2011). In Italy, compulsory education finishes at 16, but young people are entitled to receive further training (even if they access the labour market) until 18. CE courses combine classes of general disciplines, such as language and mathematics, with professional classes and laboratory traineeship. For this reason, they are generally attended either by students coming from low income families, who appreciate the shorter duration compared to classic education programmes and the vocational orientation intended to facilitate an early labour market entrance, or by students suffering from learning disabilities or previous school failures, who enjoy the inductive and practical pedagogical approach (Ragazzi 2010) and tailored pathways (Ragazzi 2008; Lauro and Ragazzi 2011). In this sense, CE training should be understood as a mixed policy, aiming at enhancing both HC and employability.

In our evaluation activity, we wanted to assess the net impact of training policies wherever possible. However, we were not able to rely on a random trial design (a random assignation to the treatment or control group). Therefore, we were forced to adopt a quasi-experimental approach, designing a control group composed of individuals who are very similar to the trainees. This is very difficult in a situation where the courses tend to be attended by less able students, who cannot be easily compared with average individuals. Our choice (Falavigna et al. 2015; Sella and Ragazzi 2016) was to rely on drop-outs and no-shows (Bell et al. 1995) as a way of checking for differences in unobservables, and indeed Heckman's (1999) tests excluded the existence of any significant selection bias (Ragazzi 2014). However, even if the theme deserves a more in-depth examination, there is a strong suspicion that the individuals of the study have different characteristics because this is a multidimensional phenomenon and the variables describing the disadvantage are often not observable and measurable.

Moreover, in some cases such as those that are the subject of our paper, it is impossible to design appropriately a control group. In the case of CE young people, they should be either at school (and in that case they cannot be compared because at the end of the CE course they are still at school) or in training. In fact, traditional counterfactual evaluation does not fit those situations where the policy necessarily covers all the eligible population.

A final element which proved unsatisfactory relates to the variable describing the employment outcome, because, again, the placement is a multidimensional phenomenon.

SEM appeared as an approach able to provide reliable descriptions of latent multidimensional variables, expressed as systems of composite indicators (CI), at the input/output/outcome levels. Previous analyses (probit models) failed in terms of describing the initial disadvantage of young and adult trainees, who generally come from a difficult family situation and social background and have a low educational attainment. In particular, family, endowment and individual network variables have not proved significant in probit models of employability (Nosvelli et al. 2012; Benati et al. 2014a, b). Moreover, SEM adopts a systemic view, able to assess the relationships between latent multidimensional variables. This is very important in the case of CE, where the solution adopted by education evaluators to assess the impact is the use of a value-added approach (Wainer 2004; Lissitz 2005; OECD 2008), which implies a comparison between entry and exit conditions.

2 The SEM Methodology and the Mixed Two-Step Approach

The construction of a CI implies the search for a suitable synthesis of a number of observed or manifest variables (MVs) in order to achieve a simple representation of a multidimensional phenomenon. Accordingly, a CI can be considered as a latent concept, not directly measurable, whose estimation can be obtained through the values of the MVs.

The existing literature offers different alternative methods in order to obtain a CI. SEM and specifically the PLS approach to SEM (PLS Path Modelling, PLS-PM) can be used to compute a system of CIs. According to this methodology, it is possible to define a CI as a multidimensional LV not measurable directly and related to its single indicators or MVs by either a reflexive or formative relationship or by both (this defines the measurement or outer model). Each CI is related to other CIs, in a systemic vision, by linear regression equations specifying the so-called structural model (or Inner Model).

The choice of using SEM as the methodological framework is useful for several reasons, particularly in that it gives: (i) the possibility of obtaining, simultaneously and coherently with the estimation method, a ranking of individuals for a specific indicator; (ii) the possibility of comparing systemic indicators in space and in time; and (iii) the possibility of estimating the hypothesized relationships without making assumptions about data distribution.

Two different approaches exist to estimate model parameters in SEM: the *Covariance-Based* (Jöreskog 1978) techniques and the *Component-Based* techniques (Wold 1982). The PLS-PM approach to SEM has been proposed as Component-Based, where the LV (i.e. CI) estimation plays a main role. As a matter of fact, the aim of *Component-Based* methods is to provide an estimate of the LVs in such a way that they are the most strongly correlated with one another (according to the path diagram structure) and the most representative of each corresponding block of MVs.

PLS-PM is a suitable tool for the investigation of a model with a high level of abstraction, in cases where the building of a system of CIs depends on different levels of construction. Higher-Order Constructs are explicit representations of multidimensional constructs that exist at a higher level of abstraction and are related to other constructs at a similar level of abstraction completely mediating the influence from or to their underlying dimensions (Chin 1998).

In Wold's (1982) original PLS-PM design, it was expected that each construct would be necessarily connected to a set of observed variables. On this basis, Lohmöller (1989) proposed a procedure to treat hierarchical constructs, the so-called hierarchical component model.

The hierarchical constructs or sayings are multidimensional constructs that involve more than one dimension and can be distinguished from the one-dimensional constructs that are characterized by a single underlying dimension. There are two main approaches existing in the literature: the Repeated Indicators Approach and the Two-Step Approach.

The Repeated Indicators Approach (Lohmöller 1989) is the most popular approach when estimating Higher-Order Constructs in PLS-PM (Wilson 2009). The procedure consists in taking the indicators of the Lower-Order Constructs and using them as the MVs of the Higher-Order LV.

The Two-Step Approach is divided into two phases. In the first step, the LV scores of the lower-order constructs are computed without the Second-Order Construct (Rajala et al. 2010). Then, in the second step, the PLS-PM analysis is performed using the computed scores as indicators of the Higher-Order Constructs.

In cases where a Higher-Order Construct is formatively related to the Lower-Order dimensions and each construct is reflectively measured by its MVs,

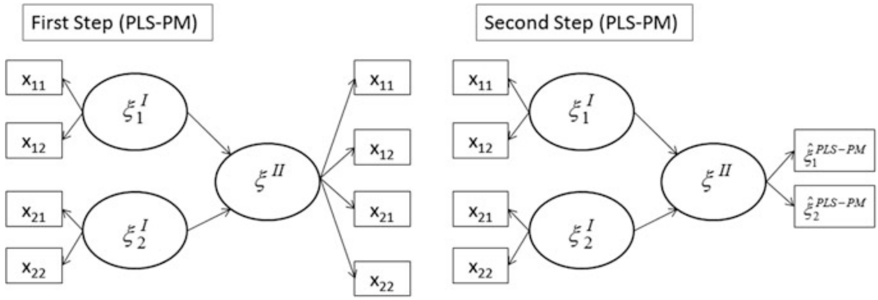


Fig. 1 Mixed Two-Step Approach implementation

the Two-Step Approach works better than the other approach. However, each approach presents some limitations. In particular, only one aspect of the Two-Step Approach is taken into account, namely the meaning of the component for each Lower-Order Construct.

In the classic Two-Step Approach, only the first component of the Lower-Order Construct is estimated without the Higher-Order Construct. This first component is the one that best represents its block of MVs. Next, these first components are included in the analysis as indicators of the Higher-Order Construct. In order to resolve the issue related to the predictive power of the component for each Lower-Order Construct, the Mixed Two-Step Approach has been developed, proposed by Cataldo (2016).

The Mixed Two-Step Approach begins with the implementation of PLS-PM as in the case of the Repeated Indicators Approach. Because the Second-Order Construct has no MVs of its own, it is considered as formed of all the MVs of the First-Order Constructs.

Starting from this structure, a PLS-PM algorithm is performed in such a way as to obtain the scores of each block. Once the scores for the blocks have been obtained, these will be the MVs of the Second-Order Construct. At this moment, once the scores of the PLS-PM have been assigned as indicators of the Second-Order Construct, the PLS-PM algorithm can be implemented (Fig. 1). Therefore, this method is proposed in order to use the component that is the best representative of its block and, at the same time, has the best predictive power on the Higher-Order LV.

3 SEM Approach to Evaluate Initial Training

This study applies the SEM methodology to the case of initial training in compulsory education in Piedmont, a region in North-West Italy. In 2013, about 4,053 young students successfully completed a multiyear Vocational Education and

Training (VET) course in that region. The SEM approach aims at evaluating the impact of the VET course in enhancing trainees' HC and, thus, employability.

3.1 SEM Approach to Evaluate Initial Training

The structural model in Fig. 2 has been developed with the aim of conceiving a policy evaluation strategy capturing any possible enhancement in individual employability due to VET. It consists of 17 LVs, describing the multidimensionality of HC and employability in the compulsory education context. In particular, HC is disentangled into both input (HCI) and output (HCO) components, in order to evaluate the direct impact of VET.

In similar studies applying the SEM approach, Dagum and colleagues (Dagum and Slottje 2000; Vittadini et al. 2003; Dagum et al. 2007) define a household HC as that multidimensional non-observable construct generated by personal ability, investments in education, and home and social environments. Partially adapting this model to the VET case study, individual HCI is a first-order hierarchical LV

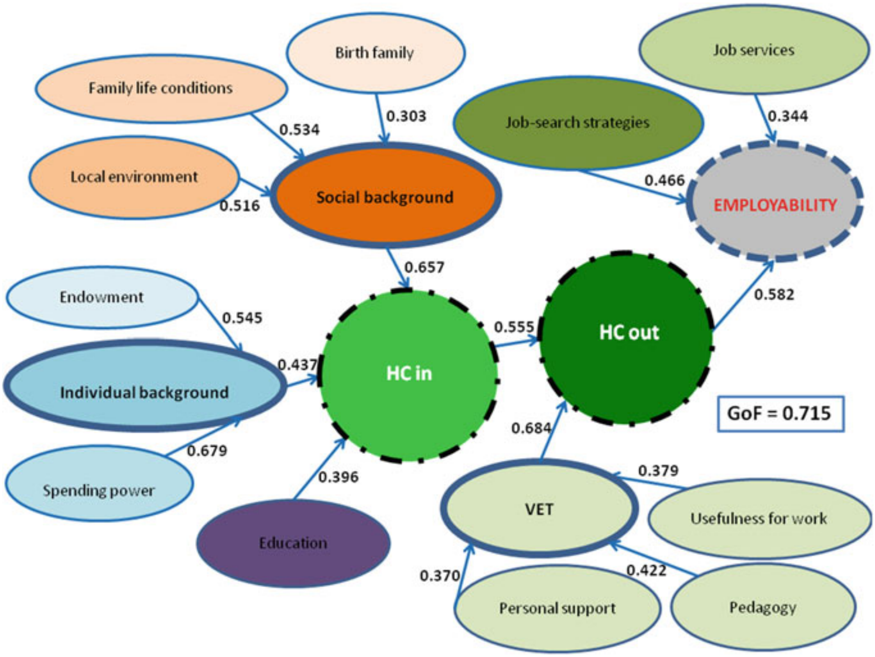


Fig. 2 Path coefficients and goodness of fit of a SEM model to evaluate the role of initial training policies on individual employability

constructed of three components: social background, individual endowment and previous education. Then, the HCI is enhanced by VET, and the resulting HCO directly influences employability, which is the hierarchical outcome, along with individual job-search strategies and the use of job services. This outcome block resembles the model of De Battisti et al. (2014), where personal employability is defined as the ability to identify and realize career opportunities and it is formed of social capital, proactivity and self-efficacy. In our case, the HCO echoes social capital, which has both professional and family components, while proactivity is described by both job-search strategies and the use of job services. In fact, there is evidence that both components enhance a trainee's probability of finding employment (Sella 2014).

Concerning measurement models, Table 1 illustrates all the MVs and their afferent first-order construct. Social background is defined by information about both the birth and present families—which are generally the same for compulsory education students—and about the neighbourhood where they live. First of all, the decision to consider the birth family block is justified by the evidence that educational disadvantage can perpetuate itself over generations, and family background can change the effects of an equal level of education (Wössmann, 2003). Hence, the corresponding measurement model describes the parents' education and employment. Secondly, the present family block reflects the effect that family characteristics have on HC investment. Following Dagum et al. (2007) and considering the lack of truly quantitative information on household wealth, the block is measured in terms of the students' perceptions about their family, economic and housing conditions. Finally, the students' opinions about their home neighbourhood (safety, cleanness, services, greenness, means of transport) are investigated as a proxy for socio-economic disadvantage, due to its role in young people's HC accumulation (Lauro and Ragazzi 2011).

Regarding the retrospective approach (Jorgenson and Fraumeni 1989; Le et al. 2005), the HC value is measured as well in terms of its production costs, demand and non-market activities. Hence, individual background is explored, in the form of personal endowment (internet, phone, means of transport) and spending power (consumption of events, sport, restaurants). Finally, educational indicators are considered (Dagum et al. 2007), that complete the HCI block and together describe the students' background at their VET enrolment.

The VET policy block, which augments the HCI, is measured in terms of the different cores of the policy, i.e. its ability to provide support for the students' needs, its promotion of school-to-work transitions, and its teaching and practical content (Benati et al. 2013).

Finally, the outer model for job-search strategies characterizes the students' efforts in different job-search channels (communication media, job referrals, employment services or contacts with previous employers), while the model measuring the students' use of job services describes how effective that particular service (vocational guidance, expert advice or mentoring) was during their job search.

Table 1 Measurement models estimates

First-Order Construct	MVs	Loadings	First-Order Construct	MVs	Loadings
Birth family	Parents' education	0.502	Personal support	Reception	0.812
	Father's job	0.685		Individual problems	0.690
	Mother's job	0.790		Learning problems	0.746
Family life conditions	Familiar situation	0.764		Personal relations	0.674
	Economic conditions	0.835	Pedagogy	Teaching	0.650
	Home	0.810		Laboratories	0.710
Local environment	Safety	0.784		Traineeship	0.703
	Cleanness	0.768		Business tutor	0.751
	Means of transport	0.692		School tutor	0.807
	Services	0.715	Usefulness for work	Labour market insertion	0.848
	Greenness	0.641		School-to-work transition	0.843
Endowment	Means of transport	0.445		Labour services	0.768
	Telephone type	0.730	Job services	Vocational guidance	0.717
	Internet connection	0.746		Expert advice	0.760
Spending power	Restaurant	0.747		Mentoring	0.743
	Shows	0.616	Job-search strategies	Means of communication	0.817
	Sports	0.615		Job referrals	0.804
	Holidays	0.514		Traineeship	0.775
Education	Grade at enrolment	0.640		Employment services	0.787
	Performance at school	0.875		Previous employers	0.755

3.2 Survey Data

The data were collected according to the computer-assisted telephonic interview methodology, surveying a representative sample of 622 young trainees who had successfully attended VET during 2013. They represent 15.3% of the total initial VET trainee intake, stratified by nationality and gender (see Ragazzi et al. 2015 for full details). The students' microdata were extracted from regional monitoring and administrative archives. The questionnaire was developed in such a way that all the MVs are ordinal Likert scales (see Table 1).

3.3 *Estimation Results*

The above theoretical model for initial training has been estimated by using the Mixed Two-Step approach to SEM, using the "plspm" package in the R statistical software (Sanchez and Trinchera 2012). All the measurement blocks are reflective (mode A), due to the algorithm requirements in repeated estimation.

Table 1 presents the loadings for each MV, i.e. the OLS coefficient of a simple regression of the MV on its LV that describes how the MV reflects the corresponding LV. The block unidimensionality, which is fundamental in reflective models, has been verified by principal component analysis, guaranteeing that all the included MVs reflect a single latent concept. The commonality indices in Table 2, that assess measurement model reliability, reveal that both personal endowment and spending power are quite poorly represented. On the contrary, the high R^2 indices show a good predictive power for all structural relations, while the redundancy indices on the endogenous blocks show a slightly lower goodness for the inner

Table 2 Inner model estimates

Higher-order blocks	Commonality	Redundancy	R^2	Path coefficient	Confidence interval
Birth family	0.448	0.195	–	0.303	[0.272;0.322]
Family life conditions	0.646	0.694	–	0.534	[0.507;0.556]
Local environment	0.521	0.641	–	0.516	[0.497;0539]
Social background	*0.523*	*0.510*	*0.976*	*0.657*	*[0.610;0.715]*
Endowment	0.429	0.537	–	0.545	[0.528;0.562]
Spending power	0.395	0.622	–	0.679	[0.647;0.713]
Individual background	*0.629*	*0.579*	*0.921*	*0.437*	*[0.388;0.477]*
Education	0.587	0.247	–	0.396	[0.537;0.430]
HC in	*0.425*	*0.408*	*0.961*	*0.555*	*[0.523;0.581]*
Personal support	0.537	0.677	–	0.370	[0.353;0.385]
Pedagogy	0.527	0.785	–	0.422	[0.407;0.446]
Usefulness for work	0.674	0.661	–	0.379	[0.363;0.404]
VET	*0.720*	*0.708*	*0.983*	*0.684*	*[0.659;0.709]*
HC out	*0.659*	*0.637*	*0.966*	*0.582*	*[0.546;0.623]*
Job services	0.549	0.340	–	0.344	[0.304;0.387]
Job-search strategies	0.621	0.334	–	0.466	[0.438;0.498]
Employability	*0.434*	*0.372*	*0.856*	–	–

model describing the HCI, and a higher goodness for the VET inner model. In any case, the overall model performs quite well (Goodness of Fit index = 0.72).

From an analysis of the path coefficients, it emerges that parents' education and job (birth family block, 0.30) play a minor role in assessing the student's social background, while both the perceived quality of family life conditions (0.53) and the local environment (0.52) play a stronger role. In turn, social background is the construct with the highest impact (0.66) on the HCI at enrolment, while the low impact of previous education (0.40) is explained in terms of both the fact that the block contains just a few MVs and that all initial training students have a standard middle school diploma. On the contrary, the VET block is very important (0.68) in determining the HCO, which is in turn the major determinant (0.58) of hierarchical individual employability, i.e. our output construct. On the contrary, the role of other active labour policies (labour services, 0.34) is quite limited, while job-search strategies show a more significant impact (0.47).

4 Conclusions

Policy evaluation is a completely new field of application (as far as we know) for SEM composite indicators. This implies that we had no possibility of following in the tracks of previous researchers in designing our model. Therefore, this exercise has involved a long and valuable learning process in the definition of the correct manifest variables. Starting from this experience, for non-statisticians (evaluators or policymakers) wanting to undertake the SEM journey, it must be pointed out that this methodology demands properly designed survey data. Our preliminary trials have shown clearly that it is very hard to adapt data collected for different purposes.

This initial work has nevertheless proved the enormous potentialities of the SEM approach also for policy evaluation:

- We have succeeded in addressing very well input multidimensionality. We have been able to characterize and measure the multidimensional latent variables impacting on HC when enrolling at vocational training (education, social background and individual background).
- From a practical perspective, this implies the possibility of developing and employing this model for individual profiling (in terms of careers guidance and policy customization).
- From a policy perspective this implies the possibility of employing the model to better design training policies, complementing them with actions aiming at countering some of the most striking aspects of disadvantage (critical area variables: high impact and low mean).
- We have been able to appreciate the added value of labour policies and, in particular, of training services, in a context where it is impossible to find a proper control group and, consequently, to assess the net impact via counterfactual methods. However, it must be pointed out that in this exercise, the

reliability of SEM in evaluating the net impact of vocational training is limited by the low quality of indicators describing the VET characteristics.

At the present stage of our research work, we have not been able to address outcome multidimensionality. The development of the outcome block, describing individual integration in the labour market, will be the object of our future research work. This will also allow a further reflection on pre- and post-treatment HC. However, we believe that this extended research study will be better performed on a different sample, relating to adult trainees, for the reason that the very young age of the trainees included in the sample adopted for this model causes huge placement difficulties, regardless of any difference in background.

In conclusion, SEM shows many potential advantages in terms of:

- measuring both pre- and post-treatment characteristics and
- evaluating differences between treated and non-treated populations.

Taking into account that it is impossible, for practical computational reasons, to use the same model for treated and non-treated individuals, we maintain that—where traditional counterfactual methods are applicable—SEM should not be seen as an alternative but rather as a complementary approach.

References

Bell, S., Orr, L., Blomquist, J., & Cain, G. (1995), *program applicants as a comparison group in evaluating training programs*. In Kalamazoo, MI, W.E. Upjohn Institute for Employment Research.

Benati, I., Ragazzi, E., & Sella, L. (2013). Valutare l'impatto della formazione professionale sull'inserimento lavorativo: Lezioni da una ricerca in Regione Piemonte. *Rassegna Italiana di Valutazione, 16*(56–57), 26–47.

Benati, I., Lamonica, V., & Ragazzi, E. (Eds.), Santanera. E., Sella, L. (2014a). *Gli esiti occupazionali delle politiche formative in Piemonte. 3° rapporto annuale di placement*. CNR-Ceris e Regione Piemonte, Torino.

Benati, I., Lamonica, V., Ragazzi, E., & Sella, L. (2017). I benefici delle valutazioni "ripetute". Evidenze da un'esperienza piemontese, *Rassegna Italiana di Valutazione*.

Benati, I., & Ragazzi, E., Santanera, E., Sella, L. (Eds.). (2014b). *Gli esiti occupazionali delle politiche formative in Piemonte. 2° rapporto annuale di placement*. CNR-Ceris e Regione Piemonte, Torino.

Nosvelli, M., Ragazzi, E., & Sella, L. (Eds.).(2012). *Gli esiti occupazionali delle politiche formative in Piemonte. 1° rapporto annuale di placement*. CNR-Ceris e Regione Piemonte, Torino.

Cataldo, R. (2016). Developments in PLS-PM for the building of a System of Composite Indicators. PhD thesis. University of Naples Federico II.

Chin, W. W. (1998). The partial least squares approach to structural equation modeling. In G. A. Marcoulides (Ed.), *Modern business research methods* (pp. 295–336). Mahwah, NJ: Lawrence Erlbaum Associates.

Dagum, C., & Slottje, D. J. (2000). A new method to estimate the level and distribution of household human capital with application. *Structural change and economic dynamics, 11*(1), 67–94.

Dagum, C., Vittadini, G., & Lovaglio, P. G. (2007). Formative indicators and effects of a causal model for household human capital with application. *Econometric Reviews, 26*(5), 579–596.

De Battisti, F., Gilardi, S., Siletti, E., & Solari, L. (2014). Employability and mental health in dismissed workers: the contribution of lay-off justice and participation in outplacement services. *Quality & Quantity, 48*(3), 1305–1323.

Falavigna, G., Ragazzi, E., & Sella L. (2015). Gender inequalities and labour integration. An integrated approach to vocational training in Piedmont. *Journal of Economic Policy*; *XXXI* (1), 97–120.

Heckman, J. L., Lalonde, R. J., & Smith, J. (1999). The economics and econometrics of active labour market programs. In Ashenfelter & Card (Eds.), *The handbook of labour economics* (Vol. 3). North-Holland, Amsterdam.

Jorgenson, D. W., & Fraumeni, B. M. (1989). Investment in education. *Educational Researcher, 18*(4), 35–44.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika, 43,* 443–477.

Lauro, C., Ragazzi, E. (Eds.) (2011). *Sussidiarietà e… istruzione e formazione professionale: Rapporto sulla sussidiarietà 2010*. Mondadori università, Milano.

Le, T., Gibson, J., Oxley, L. (2005). *Measures of human capital: A review of the literature* (No. 05/10). New Zealand Treasury.

Lohmöller, J. B. (1989). Latent variable path modeling with partial least squares. *Physica*. Verlag, Heidelberg.

Lissitz, R. W. (2005). *Value added models in education: Theory and applications*. Jam Press.

OECD. (2008). Measuring improvements in learning outcomes. OECD.

Ragazzi, E. (Ed.). (2008). *Perché nessuno si perda*. Milano: Guerini.

Ragazzi, E. (2010). Metodologie di intervento nella formazione dei giovani. In Savorana, A. (ed.), *Il liceo del lavoro. Il caso Scuola Oliver Twist*, Guerini, Milano.

Ragazzi, E. (2014). Effectiveness evaluation of training programmes for disadvantaged targets. *Procedia—Social and Behavioral Sciences, 141*(2014), 1239–1243.

Ragazzi, E., Benati, I., Lamonica, V., Sella, L., (Eds.) (2015). Gli esiti occupazionali delle politiche formative in Piemonte. 4° Rapporto di Placement. Regione Piemonte.

Ragazzi, E., & Sella, L. (2011). L'Assetto istituzionale dell'istruzione professionale in Italia. in Lauro Ragazzi.

Sanchez, G., Trinchera, L. (2012). *plspm: Partial Least Squares data analysis methods.* R package version 0.2–2. http://CRAN.R-project.org/package=plspm.

Sella, L. (2014). Enhancing vocational training effectiveness through active labour market policies. *Procedia-Social and Behavioral Sciences, 141,* 1140–1144.

Sella, L., & Ragazzi, E. (2016). Migration and work: the cohesive role of vocational training policies. *Mondi Migranti, 1*(2016), 139–160.

Vittadini, G., Dagum, C., Lovaglio, P. G., Costa, M. (2003). A method for the estimation of the distribution of human capital from sample surveys on income and wealth. In: *Proceedings of American Statistical Association, Educational Statistics Section.*

Wainer, H. (2004). Introduction to a special issue of the journal of educational and behavioral statistics on value-added assessment. *Journal of Educational and Behavioral Statistics*, *29*(1), 1–3.

Wilson, B. (2009). Using PLS to investigate interaction effects between higher order brand constructs. Esposito Vinzi, V., et al. (Eds.), *Handbook of partial least squares: concepts, methods and applications in marketing and related fields*.

Wold, H. (1982). Soft modeling: the basic design and some extensions. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction, part 2* (pp. 1–54). Amsterdam: North-Holland.

Wößmann, L. (2003). Schooling resources, educational institutions and student performance: the international evidence. *Oxford Bulletin of Economics and Statistics, 65*(2), 117–170.

Measuring the Intangibles: Testing the Human Capital Theory Against the OECD Programme for the International Assessment of Adult Competencies

Federica Cornali

Abstract The growing interest in human capital stems from the awareness that much of countries' social and economic development depends on it. In particular, the literature states that a good supply of human capital can assure individuals: (i) more employment opportunities, (ii) higher wages, and consequently, at the aggregate level, and (iii) greater growth in countries' wealth. In a number of transnational surveys (IALS 1994–1998; ALL 2003–2008; PIAAC 2011–2012), the OECD submitted a large sample of the adult population to questionnaire tests of their literacy, numeracy and capability of problem-solving. In this paper, I will first show the advantages and limits of using the skills assessed in the OEDC Survey of Adult Skills as human capital estimates. Second, using data from the most recent PIAAC survey, I will determine whether they provide more convincing support to the three assumptions of human capital theory.

Keywords Human capital theory • Human capital estimate • Surveys of adult skills

1 Human Capital: The Notion and Its Estimate

Human capital is among the longest-lived and most successful economic notions. The idea of treating the person as an asset that can be increased through investment can be traced back to 1961, in Schultz's seminal article "Investment in Human Capital". In the article's opening words, the author states that although it is obvious that people acquire useful skills and knowledge, it is not obvious that these acquisitions are a form of capital or that this capital is the product of deliberate investment. A few years later, Becker (1964) stated that human capital is equivalent

F. Cornali (✉)
University of Turin, Lungo Dora 100, 10153 Turin, Italy
e-mail: federica.cornali@unito.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_24

to a stock of one's own accumulated assets that allows people to receive revenue streams as if they were interested. According to Romer (1986), every year, part of the wealth is diverted from current consumption and invested in the production of ideas. Here, by contrast with other sectors, returns are not diminishing: unlike physical capital, human capital does not tend to lose its status as an engine of growth as a result of a massive accumulation, and thus can be a continuous and constant factor of economic development. The implications of the notion of human capital and its possible applications have given rise to a fully fledged "human capital theory" or, in other words, a coherent set of interrelated propositions that are at a high level of abstraction and from which empirical predictions can be made. The assumptions which make up this theory are as follows:

1. A high level of human capital leads to greater employability;
2. A high level of human capital leads to higher personal income;
3. A high level of human capital leads to more economic growth.

To test these hypotheses, the concept of human capital must be expressed in measurable terms. Observing the different states of human capital and assigning them to specific categories, however, is a daunting task. And trying to put a price tag on them, like any other good on the market, is even harder. Human capital is a complex property, with multiple dimensions. It is an intangible asset, comprising education, information, health, entrepreneurship, and productive and innovative skills, that is formed through investments in schooling and job training, as well as through research and development projects and informal knowledge transfers. Human capital can certainly be considered as wealth, but it is wealth of a particular type: it resides in people, and whether it is provided depends on their will (with the exception of slave labour, which still exists today).

Since calculating human capital consists of estimating of a person's ability to produce labour income, the members of the Chicago School of Economics (Becker 1964; Mincer 1974) maintained that a sufficiently robust representation of this ability could be provided by the "training flows" or, in other words, school enrolment rates, educational attainment and average years of schooling. Over the years, these considerations produced an enormous number of empirical studies on the outcomes of formal education in the labour market, in terms of employability and income, but also in terms of the economic growth of nations. Taken together, these studies have provided only partial confirmation for the three hypotheses of the theory of human capital. On the one hand, Heinrich and Hildebrand (2001) confirmed that more highly educated people throughout Europe have higher employment rates and higher incomes than the rest of the population, while Bassanini and Scarpetta (2001) estimated that the long-run effect on output of one additional year of education is about 6%. On the other hand, however, recent studies show that employability and rate of return on education are related to many factors (such as gender and age, and the country's level of economic development) and that in order to develop its potential, human capital is not sufficient in itself, but must be supported and facilitated by appropriate political, institutional and legal provisions that establish rules and institutions. Many authors (e.g. Woessmann 2000) attribute the absence at the micro- and macrolevel of robust empirical confirmation of the hypotheses underlying the theory of human capital to poor operationalization of the concept. Educational attainment represents this property empirically, but does not capture it directly. According to OECD (1998), the formal education (i) does not indicate what knowledge and skills are learned and to what extent; (ii) does not consider that skills can depreciate over time, as they may be forgotten or become obsolete; (iii) ignores the differences in outcomes that even educational institutions of the same type can produce; (iv) ignores non-formal education.

Though educational attainment is clearly a proxy of human capital, there are recurrent references to it in the scientific literature as a "measurement". Such respected scholars specializing in the subject as Mulligan and Sala-i-Martin (2000) have entitled their article "Measuring aggregate human capital". This expression is misleading. Strictly speaking, it would be possible to "measure" human capital only if we have some unit of measurement or standard quantity that can be used as a benchmark. A measure of this kind obviously does not exist. It would be a pity, however, to abandon all hope of achieving some sort of "quantitative evaluation" of a property as significant as human capital. Without such a quantitative measure, the stock of human capital available in different areas would remain unknown, and the effects arising from it could not be weighed. Given that knowledge and skills, like many other personal characteristics, can be described as continuous properties, applying appropriate scaling techniques can provide an acceptable quantification of human capital. Making direct measurements of human capital means subjecting samples of the population to standardized stimuli, which elicit responses that can be evaluated and interpreted in accordance with specific criteria and make it possible to draw up significant competence profiles.

2 Testing the Theory of Human Capital against Surveys of Adult Skills

An attempt to making direct measurements of human capital has been realized by international programme for directly assessing the so-called key information-processing competencies launched by the OECD. This programme—consisting of three surveys: IALS (International Adult Literacy Survey); ALL (Adult Literacy and Life Skills Survey); and PIAAC (Programme for International Assessment of Adult Competencies)—started in the early 1990s and investigated the different components of human capital submitting the respondents to the proficiency tests.

The most recent PIAAC, surveyed around 166,000 adults aged 16–65, were in 24 countries from 2011 to 2012. This survey directly assessed the skills of the adult population in three domains deemed fundamental to effective and successful participation in the economical and social life of advanced economies, namely: literacy, numeracy and problem-solving in technology-rich environments. To make the best possible use of the enormous body of data offered by the PIAAC survey, in this work I have decided to reduce the number of variables representing the different domains of literacy by summarizing them in a single index, which we can consider sufficiently representative of the individual's human capital. This index—hereafter, the « HC PIAAC Index »—was obtained by carrying out principal component analysis on the scores describing each respondent's literacy, numeracy and problem-solving. The analysis excluded France, Spain and Italy because these countries have not given evidence relating to problem-solving. The « HC PIAAC Index » has a score that ranges from −3.91 to 3.69. The decision to use a single index was validated by analysis results. In fact, the first extracted component has a very high eigenvalue (17.03) and can faithfully represent all the variables, accounting for 85.1% of the variance. The OECD researchers state that although they chose to represent literacy with three separate scales, they are highly correlated and, consequently, "a robust general literacy factor can be found in the respondents" (Rock 1998, p. 142).

Coulombe, Tremblay and Marchand propose using results of adult literacy skill surveys to "finally harmonize statistics" on human capital (Coulombe et al. 2004, p. 9). This opinion is shared by many authors, and the OECD suggests that its "survey of adult skills should offer a more accurate picture of skills relevant to the labour market and could help to explain differences in earnings and economic growth" (2013, p. 104). Proceeding as these authoritative sources suggest, I tested the three hypotheses of the theory of human capital using the « HC PIAAC Index », rather than the usual formal education indicators. I thus attempted to determine whether a good stock of human capital guarantees more employment opportunities and higher wages for individuals and, at an aggregate level, promotes the growth of countries' economy. As regards the extent to which a better stock of human capital corresponds to a higher probability of having a job, the logistic regression analysis indicated that the relationship between the « HC PIAAC Index » and the employment status of the survey respondents is not relevant ($R^2_{\text{Nagelkerke}} = 0.015$). To refine the analysis and protect one of the core notions of human capital theory from falsification, I tested numerous models (not listed here for brevity), by controlling some representative variables of specific personal condition or general condition that the literature indicates as particularly significant in determining the employability (see Sect. 1). The logistic regression model, including the variables « HC PIAAC Index », age, gender and level of economic development of the country (in terms of gross domestic product), reproduces less than 10% of the observed variance ($R^2_{\text{Nagelkerke}} = 0.092$). Furthermore, the scores of the variables in the equation (Table 1) show that the greater relevant factors for occupational status are primarily the high level of economic development of the country of residence, age and, lastly, human capital.

Also, the second hypothesis is not supported by the PIAAC data, and in fact, they do not show no effect of human capital on individual earned income. Moreover, testing several multiple regression models (even in this case not listed for brevity) did not reveal any causal relationship, due to spurious effects and multicollinearity of variables considered. The results of the hierarchical regression

Table 1 Logistic regression model for variables affecting “Employed” (N = 50,722)

Variables in the equation							
		T	E. S.	Wald	gl	Sign.	Exp (B)
Step 1	Human capital	−0.397	0.020	377.019	1	0.000	0.672
	Age	−0.057	0.002	1189.727	1	0.000	0.945
	Gender	0.039	0.036	1.137	1	0.286	1.040
	Economic development	0.057	0.009	41.183	1	0.000	1.058
	(Costant)	−0.760	0.084	81.508	1	0.000	0.468

Source Author’s calculation from PIAAC (2011–2012) dataset

Table 2 Hierarchical regression analysis for variables affecting “Monthly earnings PPP corrected $US” (N = 47,916)

Coefficients					
Model 2	Unstandardized coefficients		Standardized coefficients	t	Sign.
	T	Std. error	Beta		
(Costant)	3442.40	1622.75		2.121	0.034
Age	9.53	31.58	0.002	0.302	0.763
Gender	704.40	451.63	0.008	1.560	0.119
Economic development	−332.33	253.49	−0.007	−1.311	0.190
Human capital	−598.06	498.24	−0.007	−1.200	0.230

Source Author’s calculation from PIAAC (2011–2012) dataset

model, including the variables « HC PIAAC Index », age, gender and level of economic development of the country, are not relevant as explanation model (R^2 = 0.000) having standardized coefficients near zero (Table 2).

The fact that the second hypothesis of the theory of human capital also failed to be confirmed is probably because it is based primarily on the assumption two conditions. The first supposes that workers’ earnings are entirely determined by their marginal productivity. The second assumes that the possession of a certain skill is offset by improved earnings, encouraging people to receive training to acquire it. Nevertheless, more differences than similarities emerge from a comparison of the labour markets to this ideal type. First of all, many factors other than productivity determine wages, including changes in the economic and production system, the power relations between different distributional coalitions (trade unions, business associations), the action of various forms of discrimination (ethnic, gender) and, not least, the size of the company (large companies generally offer better earning opportunities). Second, a portion of earnings is also determined endogenously through individual choices based on attraction or repulsion towards certain types of employment. Those who prefer a secure job, or need to work part-time or avoid shift work may accept lower pay provided that these needs are met. Again, employees often fail to find a job that will allow them to express their full potential. The perfectly competitive market is the simplest representation of rational behaviour, where individual choices are made in complete independence, regardless of each individual's history, and where each actor has a perfect knowledge of the market. In reality, this is not the case. People operate under uncertain conditions: They do not know all the options available to them, and even if they did, they would not always be able to predict the consequences. The context of the choice is not clear, but susceptible to interpretations that are often in conflict with each other. If in an ideal situation the return on a skill is commensurate with the investment in education needed to acquire it, in the contingent world, returns are much more volatile because of a number of phenomena, some long familiar, others more recent. The former include overeducation—with the consequent competition among individuals to get better jobs. The latter include early obsolescence of knowledge, which calls for constant effort and adaptation to changes in labour market demand. Lastly, the direct measurements of human capital also fail to corroborate the third hypothesis, i.e. that a large stock of human capital leads to more economic growth. For over forty years, since technological change has been identified as a fundamental determinant of total factor productivity, externalities due to human capital have been regarded as

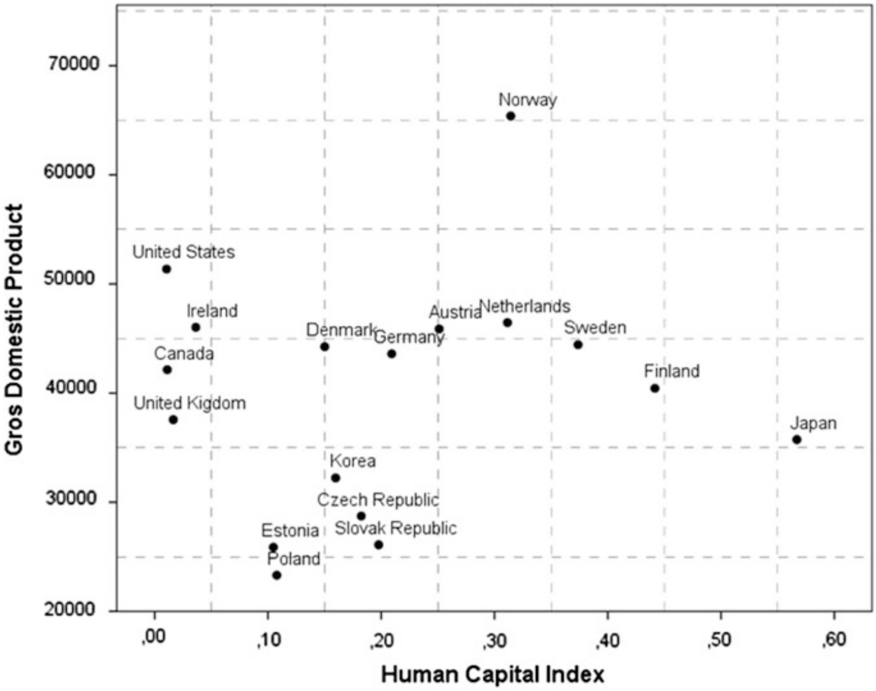


Fig. 1 Relationship between gross domestic product and human capital in several OECD countries. *Source* Author's calculation from OECD and PIAAC (2011–2012) datasets

essential to growth in the long run. This connection, however, although there is some positive evidence (see Sect. 1), is not unequivocal. This lack of corroboration is clear from the figure (Fig. 1) which compares the « HC PIAAC Index » and gross domestic product in selected countries. The densification of the countries in a shape almost in X rather shows the presence of a latent factor, not detected, which correlates positively with a variable and negatively with the other.

More generally, the data collected does not rule out the possibility of a reverse causal relationship: higher levels of education and skills in a population could be the result of achieving a good level of economic growth and not vice versa. The many deviant cases and the failure to find specific generative mechanisms lead us to conclude that there is still no solid theory correlating human capital to countries' economic growth.

This rapid examination has shed light on the limits and potential of the notion of human capital. Dissatisfaction with its representation in purely economic terms has stimulated renewed attention to how it should be defined. The original concept of *basic human capital* (BHC) has thus been joined by that of *wider human capital* (WHC). Unlike BHC, which typically includes the set of features directly related to individual productivity, WHC includes all the "knowledge, skills, competencies and attributes expressed by individuals that facilitate the creation of personal, social and economic well-being" (OECD 2001, p. 19). Introducing the concept of WHC, however, does not seem to have been decisive: the areas referred are highly dissimilar—ranging from innate qualities to education, and to physical and psychological health. In sociology, the usefulness of a concept is defined by its information capacity and its ability to stimulate more satisfactory explanations. Regarding the first requirement, WHC is undeniably effective: it provides an accurate, comprehensive representation of human capital. By contrast, the heuristic potential of the concept of WHC seems rather less obvious. The variety of connotations it covers is so extensive and so miscellaneous that the theory built around it is useless. Its rich hypothetical framework—in addition to individual earnings and economic growth, WHC also takes social cohesion, democratic participation and many other aspects into account—is countered by the excessive number of causal factors. Ultimately, this results in a kind of tautology in which "everything" explains "everything".

3 Conclusions: A Theory without Empirical Evidence?

Inaccurate operationalization of the concept, inconclusive evidence, a circular explanatory process, a certain abuse of the *ceteris paribus* clause: with all these failings, is the empirical theory of human capital fragile and imperfect? While Blaug (1976) admits that the theory of human capital is "poorly substantiated", McCloskey (1983) hands down a decidedly more trenchant verdict, pointing out that the weakness of its causal links relegates it to the status of mere pseudo-explanation, "a metaphor" typical of an unscientific, rhetorical way of understanding. According to several authors, the theory of human capital, although fruitful—it is an important starting point for studies in many different areas (wage differences, economic growth and development economics)—is based on inadequate assumptions. Nevertheless, the many reservations that have been expressed do not seem to have provided good grounds for abandoning the theory of human capital. Indeed, there has been as much consensus as there has been criticism. The limits of human capital theory belie its undeniable success. To resolve this contradiction, Poulain suggests a phenomenological approach, which "does not consider human capital as a conceptual construction, but rather a category of practice, a spontaneous interpretation" (2001, p. 92). This does not seem to be a useful suggestion, but rather an invitation to retreat from disciplined study. A call for more rigorous conceptualizations and operationalizations would be far more helpful.

References

Bassanini, A., & Scarpetta, S. (2001). *Does human capital matter for growth in OECD countries? Evidence from pooled mean-group estimates*. OECD economics department working papers, 282, OECD Publishing.

Becker, G. (1964). *Human capital. A theoretical and empirical analysis, with special reference to education*. New York: Columbia University Press.

Blaug, M. (1976). The empirical status of human capital theory: A slightly jaundiced survey. *Journal of Economic Literature*, *XIV*(3), 827–855.

Coulombe, S., Tremblay, J.-F., & Marchand, S. (2004). Literacy scores, human capital and growth across fourteen OECD countries. In *Statistics Canada catalogue*, 89–552-MIE200411.

Heinrich, G., & Hildebrand, V. (2001). *Public and private returns to education in the European Union—An appraisal*. Luxembourg: European Investment Bank.

McCloskey, D. (1983). The rhetoric of economic. *Journal of Economic Literature, 21*(6), 481–517.

Mincer, J. (1974). *Schooling, earnings and experience*. New York: Columbia University Press.

Mulligan, C. B., & Sala-i-Martin, X. (2000). Measuring aggregate human capital. In *NBER working paper* (p. 5016).

OECD. (1998). *Human capital investment. An international comparison*. France: OECD Publishing.

OECD. (2001). *The well-being of nations. The role of human and social capital*. France: OECD Publishing.

OECD. (2013). *The survey of adult skills: Reader's companion*. OECD Publishing.

Poulain, E. (2001). Le capital humain, d'une conception substantielle à un modèle représentationnel. *Revue économique, 52*(1), 91–116.

Rock, D. (1998). Validity generalization of the assessment across countries. In S. Murray, I. S. Kirsch, L. Jenkins (Eds.), *Adult literacy in OECD countries. Technical report on the first international adult literacy survey* (pp. 135–142). Washington, U.S.: Department of Education.

Romer, P. M. (1986). Increasing returns and long run growth. *Journal of Political Economy, 94*(5), 1002–1037.

Schultz, T. W. (1961). Investment in human capital. *American Economic Review, 1*(2), 1–17.

Woessmann, L. (2000). Specifying human capital: A review, some extensions and development effects. In *Kiel working paper* (p. 1007).

Analysis of the Employment Transitions and Analysis of the Unemployment Risk in the Social Security Account Statements of the Patronato ACLI

D. Catania, A. Serini and G. Zucca

Abstract The reforms of the labor market and the social security system triggered deep changes in employment and pension careers of Italian citizens, characterized by a greater job insecurity and less pension prospects consistency. The study describes the labor changes through the social security account statements of more than 95 thousand users who between 2008 and 2014 have turned to the Patronato ACLI. In particular, it was realized a study of the main career transitions through the use of social security account statements in the Patronato ACLI database and a survival analysis in case of unemployment risk. The results of the analysis made it possible to seize the main employment trends which characterized the labor market since 1960 to today. The originality of this study was to observe the occupational changes (flexibility of the labor market and worker mobility) from a "micro" perspective, starting from the "pension biography" of workers registered in the INPS fund. This approach has been developed by the Riccardo Revelli laboratory.

Keywords Public archives · Social security account statements · Employment transitions · Survival analysis · Unemployment risk

1 Studying Pensions with Administrative Data: Public Archives Capability and Limits

Recently, the study released by INPS on the future of pensions in Italy, significantly titled "Not for cash, but for fairness," brought to the attention of the mass-media and more generally of the public debate, the issue of future sustainability of the

D. Catania (✉) · A. Serini · G. Zucca
IREF, Via Marcora 18/20, 00153 Rome, Italy
e-mail: danilo.catania@acli.it

A. Serini
e-mail: alessandro.serini@acli.it

G. Zucca
e-mail: gianfranco.zucca@acli.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_25

Italian pension system. The work overseen by the President of the Institute, Tito Boeri proposes elements of great interest from a methodological point of view, as well as being interesting in terms of substance that is for the proposal to revise some of the basic institutions of the pension system by offering greater protection to vulnerable groups such as the unemployed over-50s.

In particular, the SIA55, the last instance income that protects from the risk of poverty people with little opportunity to re-employment was processed through a microsimulation model based on the integration of administrative data and survey. As stated in the methodological note of the report (INPS 2015:47)

> This model is based on a representative sample of the Italian population given by the survey ISTAT IT-SILC for the year 2012. […] For analysis, all income information, originally present in the survey IT-SILC, have been replaced through the tax data available to INPS, so as to minimize typical errors on income surveys and thus ensure the adequate strength to the estimates on the costs (and the effects) of the measure in object.

The concatenation between survey data and administrative data allows to overcome quality problems as arising from the reluctance of respondents. In the case of the INPS report, the income figures, typically suffering from underestimation phenomena, are "corrected" with "objective" fiscal data, allowing it to produce more accurate and robust estimates on the cost of the social security measure simulated. This is just one possible use of administrative sources for the evaluation of social policies. In general, at the international level, we are seeing a strong integration between primary and secondary data, either because the information needs are more extensive and can not all be covered with direct measurements; either because the primary data have a higher production cost. In this respect is exemplary the choice, already practiced in other major countries, to create the next population census using a mixed approach, through a combination of administrative and primary data. The use of so-called process data, any information resulting from administrative or management procedures, is a significant tendency in official statistics. In Italy, one of the most significant experiences is the cooperation between Istat, Ministry of Labour and INPS: through the integration of data bases of various origins, the three institutions have been publishing for four consecutive years (last edition 2013) the Report on Social Cohesion, extensive database, tables, and ad hoc calculations based on socio-economic issues.

These examples of institutional collaboration are not the only way to exploit administrative data. Another example that highlights a different way of using public sources is the archive of work histories distributed by INPS through LABORatorio Riccardo Revelli of Turin University. WHIP (Work Histories Italian Panel) is a database of individual work histories, built by INPS archives data. The target population consists of all people—Italian and foreign—who have carried out part or all of their working career in Italy. A large representative sample was extracted from it.[1] For each of these persons, they are observed the main episodes that

[1] In the standard file the sampling ratio is about 1: 180, for a dynamic population of about 370,000 people (these figures will be doubled for the full version).

characterize their working career. The complete list includes the position of an employee, the "para-work" periods, the self-employed as a craftsman or a trader and, for certain professional activities, retirement, as well as periods in which the individual has benefited from social benefits—such as unemployment benefits or mobility allowances. At the moment, the WHIP database covers the 1985–2004 period.[2]

WHIP is an invaluable source for reconstructing the medium-term trajectories of the Italian labor market, since it allows the development of longitudinal analysis, keeping track of the occupational trajectories of individuals. In the notes to the release of the latest version of the database, however, Leombruni et al. (2010: 3) highlight a number of problems of quality of the data: the first is the correct identification of the target population because the management archives of INPS offer a fully representative for "all and only people over the age of 15 years […] who have had a job as private employees in industry, construction and services; as holders of craftsmanship activities; as traders; as quasi-employees."

The second issue concerns the recording errors of tax codes, since the episodes of a working individual may submit not matching codes with the identifier code in the INPS anagraphic archive: the errors in the compilation of personal data are frequent and require to be corrected with complex control procedures and remediation of the database. In general, the indications on the use of management data INPS show that in the face of a clear benefit in terms of empirical research, there are non-secondary problems of data quality, especially since it regards databases with millions of records. The most evident consequence is the amount of time that elapses between the data reference period and the release of the matrix for use in research. The problem could be overcome, or at least reduced, by providing improvements upstream of the production of the data (automatic controls at the moment the process of information), or at an intermediate stage, for example, by mandating the general statistical actuarial Coordination of INPS to take care of the cleaning process. Overall, both the solutions reveal that is still incomplete, at present, the institutional awareness of administrative data utility for purposes of socio-economic research.

In the wake of these experiences, this paper presents an experimental use of social security data with empirical research objectives. The data source is not institutional, however, as the work presented in the following pages is based on an "ad hoc" extraction by the Patronato ACLI administrative archives.[3] As it is known,

[2] From the other point of view, the work episodes that are not registered in WHIP are those from the public sector or as freelancers with an autonomous security fund. The WHIP subsection on employment is a Linked Employer-Employee Database: in addition to data on the employment relationship, data concerning the firm in which the person is employed are also present, thanks to a combination with the Observatory of the INPS companies (Leombruni et al. 2008).

[3] This paper is a summary of the analytical work that has seen the participation of Iref researchers and professionals of the Patronato ACLI (Franco Bertin, Claudio Piersanti, Nicola Preti and Antonio Turrini) with the coordination and supervision of Giuseppe Foresti, Research Office responsible of the Patronato ACLI, who gave the major contribution to the analysis and

the Patronati are institutions of social assistance non-profit organization, which perform functions of representation and protection in favor of the workers (employees and self-employed workers), pensioners and all citizens on the Italian country.[4] The assistance and advice of a patronage is aimed at achieving social security, health care and social welfare nature, including those relating to emigration and immigration. Beyond the functions, the important element is that the Patronati have the right to have access to the databases of social security services, through the assisted person authorization. This possibility offers to the research institutions the opportunity to work on comparatively better data than the ones provided by the INPS, as one of their main activities is the so-called pension calculation. On entry into retirement, many workers are turning to the Patronati to present the request to retire; in order to be able to formalize the practice, operators examine the user's social security account situation and, if they found inconsistencies, they report the situation to the INPS, thus obtaining a "clean" bank statement. This service gives rise to a better "internal" database compared to the INPS one. How related by the Patronato ACLI operators consulted for the preparation of the study, not always INPS corrects the Patronati reports. To be precise, the remediation of the social security accounts, despite being valid for obtaining social security benefits, not necessarily leads to a correction in the administrative INPS archives. In other words, the Patronati archives and the INPS archives are not perfectly aligned. This does not mean that the archives managed by the Patronati are free from data quality problems. It would be interesting to have comparative studies to provide comparisons on specific database segments, because as you can imagine the temporal stratification is one of the factors that influences the quality of data; by switching to fully computerized procedures, material errors decreased; however, it is obvious that for subjects with a social security history that began decades ago, when the procedures were in large part manual, the possibility of inconsistencies is much higher.

The Patronati archives are extremely under-used because, as we will verify below, they present a non-functional structure for research purposes. In particular, the biggest problem is the absence of a data matrix structure, for which are necessary operations of data shift, activities that still can be realized through relatively simple procedures. Beyond the technical and practical aspects, the cognitive value of the archives available to the Patronati is considerable. We hope that this first trial could be an example to increase the knowledge on the future of the social security system.

(Footnote 3 continued)

interpretation of data. Special thanks to Paola Vacchina (President of Iref), who promoted this analysis work when she was president of the Patronato ACLI.

[4] The Patronato institutes were recognized by the italian State with a Legislative Decree CPS, July 29th, 1947, n. 804, which already contained the first regulations governing and regulating these institutions. Then the law of March 30th, 2001 n. 152 (New legislation for institutes of patronage and social care, published in the Official Gazette no. 97 of April 27th, 2001) launched a reform that re-evaluates the role and redefines the tasks of the Patronati.

In the next few paragraphs, it will illustrate the procedures adopted for the construction of the contributive transitions matrix (in Sect. 2.1) and the main results of the contributive transitions analysis and of the occupational survival (in Sect. 2.2). Finally, the conclusions will return to the issue of quality of data, some considerations that derive from the experience with the Patronato ACLI.

2 A First Experimentation on Social Security Account Statements

The basic orientations that guided our investigation are due to the attempt to read the changes in the labor market, intervened from the postwar to the present days, through the heritage of information and knowledge that the ACLI Patronage matured during his 70 years of activity.

Specifically, we set an objective to explore the informative potential in more than 95 thousand pension accounts selected from the Patronato ACLI database. An exploration conducted with the longitudinal analysis typical tools, through the contributory careers of the Patronato users, which made it possible to grasp the changes in the working world. It is, therefore, of a cross section of the social security sector that does not pretend to be exhaustive, but represents the Acli point of view. A point of view which, though partial, allowed to focus on the salient features of the metamorphosis taking place in the labor and social security system Italian (insecurity, procreative deferral, short circuit of the generational pact, etc.). Not only that, the longitudinal analysis gave way to deepen the functioning of the contributive accounts internal to the different social security funds.

2.1 Construction of the Transition Matrix

The information corpus from which it was extracted the data base is constituted by more than 388 thousands INPS social security accounts, stored, from 2008 to 2014, in the Patronato ACLI archive. It is a repository organized on the basis of the institutional and administrative tasks that the Patronato plays for workers and public institutions—INPS and Ministry of Labour. More specifically, the Patronato operators primarily examine the archive INPS to test the regular payment of contributions and to calculate the date of retirement of users who turn to local offices. The operator can then view the account statement present in the INPS database and even download a copy. However, Patronato operators download, from the INPS archive, statements not subjected to tests, and eventual corrections made by operators are executed in offline mode. It is not unusual, therefore, to find missing data, inconsistencies, and errors in the INPS statements. In addition to that, the INPS archive has not a unique system of encoding information. The restructuring of the

INPS database that have occurred over the years produced a layering of coding systems because of the overlap of the new archives to the old ones. This created a layered coding system: the same information shall be indicated with different codes according to the contribution period to which it relates.[5]

The extraction of the transition matrix from the corpus of the account statements of the Patronato ACLI took place in three operational steps:

1. re-organization of the corpus of the account statements for statistical purposes and selection of analysis variables;
2. construction of the analytical basis through the selection of records without errors and missing values;
3. normalization of the coding system by assigning a unique code to the different contribution states.

To make a "cases for variables" matrix, it was made a transposition of the table of social security account statements.

The transposition procedure of the pension account statements generated such a number of columns to be unmanageable in the stage of data processing. It was contained the number of columns of the matrix by imposing a dual criterion of simplification: selection of a subset of variables and definition of a maximum limit of contribution periods to be analyzed.

The selected variables were ten: three on the anagraphic status of workers (sex, age, and marital status) and seven related to their social security career. The maximum number of contributory events was determined with a test matrix with which it was tested the capacity (hardware and software) of the computers to process large data masses. At the end of the load test, we have set the threshold in ninety contributory events, thus bringing to 634 the number of the variables (columns) of the matrix of transitions.

Parallel to the procedures followed in determining the size of the matrix, checks were carried out on the quality of data, with the application of logical filters, consistency checks and integrity. Furthermore, it proceeded to the normalization of the INPS data coding system, by assigning a unique contributory code to each contributory situation.

At the end of the data matrix preparation procedures, from 388 thousand social security account statements stored in the Patronato ACLI database, 95,646 suitable records were selected for the analysis of employment transitions. It is a sample of taxpayers equally distributed by gender: 49.3% women and 50.7% men (see Table 1). In contrast, the age distribution (see Table 2) shows a concentration of taxpayers in the age close to retirement age: 51–60 years (32.1%) and 61–65 years (36.2%). In a manner consistent with the distribution by age of the sample, the majority of taxpayers began to pay social security contributions in the seventies

[5] For example, the employee's pension status is indicated with four codes (100, 103, 106, and 108 and 198).

Table 1 Sex

Sex	N	%
Donne	47.180	49.3
Uomini	48.446	50.7
Totale	95.626	100.0

Source IREF processing of Patronato ACLI data

Table 2 Age

Age	N	%
18–30	1.950	2.0
31–40	2.872	3.0
41–50	5.332	5.6
51–60	30.742	32.1
61–65	34.611	36.2
66–70	16.185	16.9
71–80	3.872	4.0
81–90	82	0.1
Total	95.646	100.0

Source IREF processing of Patronato ACLI data

Table 3 Period and mean age of the users' contribution start

Users' contribution start	N	%	Mean age of contribution start
Before 1960	1.326	1,4	16
1960	22.706	24,4	17
1970	43.017	46,2	20
1980	10.660	11,4	26
After 1980	15.492	16,6	38
Total	93.201	100,0	23
Missing data	2.445	–	–

Source IREF processing of Patronato ACLI data

(46.2%—see Table 3) and, to a lesser extent, in the sixties (24.4%), periods in which they entered the pension system at 20 years of age.

In summary, the extracted sample reflects the typical user of the Patronati: workers close to retirement, who turn to a Patronato structure for services related to retirement: checking the contributory situation, calculating the date of the retirement, starting the retirement practice.

In the next section, we describe the main results of the longitudinal analysis on employment careers and on social security accounts of the taxpayers.

2.2 Contributory Transition and Unemployment Risk Analysis

The contributory transitions were processed by the variable named “security funds type.” In the social security account statements, for each contribution period, it shows the social security fund in which the payment was made. The security funds indicated in the social security account are six: employees, farmers, craftsmen, traders, farmers (CDCM), and professionals/not regular employees (specific contributive fund called “gestione separata”—GS). It has also been defined the “Unemployed” mode to indicate the condition of contribution absence, due, in most cases, to a state of unemployment.

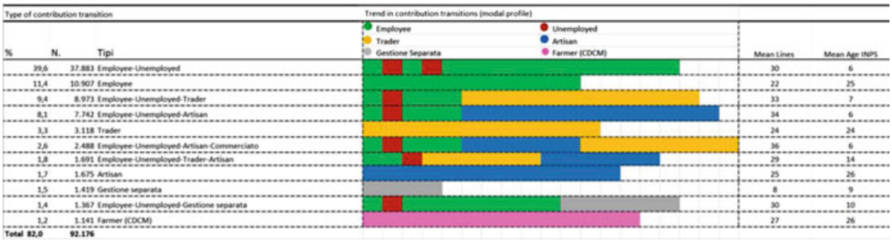
In the course of working life, a worker can progress in more than one contributory state or stay in the same state. In theory, in the social security account statements, there may be a large number of combinations, defined within a contributory variability continuum whose poles are: the persistence throughout the contribution period in the same pension category (contributory uniformity); or to pass for all contribution states (maximum contributive heterogeneity). Within these opposites take shape the contributive trajectories/transitions of users of the Patronato ACLI.[6]

Table 4 shows the most frequent transitions, ordered descending. The length of the transitions is based on the average value of the rows of the social security account statements. The condition of employee is a trait common to most contributory transitions. Often the contributory career of the population under investigation begins and ends in a state of employee: 51% of the sample has in effect a contributory career of employee; 39.6% employee interspersed with periods of inactivity and 11.4% employee without any interruption of contribution. In other cases, the contributory transitions take booting to an employee condition and then move on to other social security funds, especially those of traders and craftsmen: Employee-Unemployed-Trader, 9.4% and Employee-Unemployed-Artisan, 8.1%.

The characterization of the transitions for the starting contribution period (in Table 5) confirms the importance of the status of employee in the formation of pension contribution transitions. In all the pension contribution periods, the first

[6] The identification of the transitions occurred through the counting of different categories drawn from the social security course of each user. For each category, it was counted as many times as it appeared in the social security account statements, for example: if an extract is made up of four years in which it appears, in each year, the state “EMPLOYEE”, the relationship within the category “EMPLOYEE” amounts to 1, while in the other categories the value is null. In this case, the transition is defined of only employee; conversely, if in the account statement there is an alternation between employment and inactivity (“unemployed-dependent-unemployed”) the value, both in the category “EMPLOYEE” and in the category “UNEMPLOYED”, is 0.5, while in the other categories is null. In this situation, the transition is of “Employee-Unemployed” type. Proceeding this way we have identified 62 types of transitions.

Table 4 Contributive transitions *(*To place the sections in which the contributive trajectory is interrupted by an event of inactivity, such as in the case of employee-unemployed, it was used a prevalence criterion linked to the modal value of the unemployment distribution for that type of transition. Chart. 1 shows the trend of the unemployment incidence over the years. In the employee-unemployment transition, the periods of inactivity tend to be more frequent at the beginning of the contributory career, and they decrease with the increase of years of work. In particular, the highest unemployment incidence is present in the second (47.2%) and fourth year (27.6%). In these, two periods are recorded, therefore, the highest values of the unemployment. Therefore, consistent with our prevalence criterion in the representation of "employee-unemployed" transition, the unemployment periods were placed early on, near the second and fourth years of the contributory career. A similar procedure was used in the construction of other contributory transitions.*)*


Trend in contribution transitions (modal profile)
Employee
Trader
Gestione Separata
Unemployed
Artisan
Farmer (CDCM)

Type of contribution transition				
%	N.	Tipi	Mean Lines	Mean Age INPS
39,6	37.883	Employee-Unemployed	30	6
11,4	10.907	Employee	22	25
9,4	8.973	Employee-Unemployed-Trader	33	7
8,1	7.742	Employee-Unemployed-Artisan	34	6
3,3	3.118	Trader	24	24
2,6	2.488	Employee-Unemployed-Artisan-Commerciato	36	6
1,8	1.691	Employee-Unemployed-Trader-Artisan	29	14
1,7	1.675	Artisan	25	26
1,5	1.419	Gestione separata	8	9
1,4	1.367	Employee-Unemployed-Gestione separata	30	10
1,2	1.141	Farmer (CDCM)	27	26
Total 82,0	92.176			

transition is Employee-Unemployed, with frequencies ranging from 43.5% of the sixties to 54.2% of the subsequent periods of the eighties. However, in the first and last period (1960s and after the 1980s), we note deviations from the model based on employee category, with a significant presence of transitions organized on CDCM categories (the 1960s) and the specific contributive fund called "gestione separata" (to the final period). These differences are a reflection of laws that have extended the insurance and pension protections for special categories of workers: in effect, at the turn of the fifties and sixties, the social security protections were extended to farmers (in act n. 1047, 1957 and n. 9, 1963) and in the nineties, with the so-called Dini reform (act no. 335 of 1995), to free professions and atypical workers.

In this brief description of the analysis of transactions, we can already glimpse the potential of the techniques and longitudinal patterns in the knowledge of social phenomena. In the specific case of the Patronato ACLI, the analysis of transitions has opened new areas of knowledge both in a management-specific point of view (linked to the development of more flexible social security services to different contributive profiles) and in a study perspective, by offering food for thought on the labor market changes.

The framework of knowledge and ideas that emerged from the analysis of the transitions has been enriched with the results of the survival analysis (Fleming and Harrington 1991; Hosmer and Lemeshow 2008; Cleves 2010). Specifically, it was made a survival analysis to the first occupation, considering a 10-year observation period. The analytical model is based on the formation of two groups: the "survivors," those who have kept their first job for the entire reference period; and the

Table 5 Basic transitions for starting contribution period

Rank	Transitions	1960s	%	Transitions	1970s	%
1	Employee-Unemployed	7.704	43,5	Employee-Unemployed	17.900	48,8
2	Employee-Unemployed-Artisan	3.168	17,9	Employee	4.586	12,5
3	Employee-Unemployed-Trader	3.088	17,4	Employee-Unemployed-Trader	4.440	12,1
4	Employee-Unemployed-Artisan-Trader	1.124	6,3	Employee-Unemployed-Artisan	3.452	9,4
5	Employee	496	2,8	Trader	1.140	3,1
6	Employee-Unemployed-GS	445	2,5	Employee-Unemployed-Artisan-Trader	1.120	3,1
7	Farmer-Unemployed-Employee	380	2,1	Trader-Unemployed	891	2,4
8	Trader-Unemployed	288	1,6	Artisan	749	2,0
9	Farmer	227	1,3	Farmer	572	1,6
10	Farmer-Unemployed	196	1,1	Employee-Unemployed-GS	517	1,4
Total		17.703			36.688	
Rank	Transitions	1980s	%	Transitions	After 1980s	%
1	Employee-Unemployed	3.909	41,9	Employee-Unemployed	8.067	54,2
2	Employee	1.464	15,7	Employee	2.631	17,7
3	Employee-Unemployed-Trader	923	9,9	GS	1.127	7,6
4	Trader	890	9,5	Trader	781	5,2
5	Employee-Unemployed-Artisan	663	7,1	GS-Unemployed	517	3,5
6	Artisan	396	4,2	Employee-Unemployed-Trader	374	2,5
7	Trader-Unemployed	359	3,8	Artisan	305	2,0
8	Artisan-Unemployed	150	1,6	Employee-Unemployed-Artisan	254	1,7
9	Farmer	148	1,6	Employee-Unemployed-GS	240	1,6
10	Employee-Unemployed-Artisan-Trader	148	1,6	Employee-Unemployed-Agricoltura	164	1,1
Total		9.339			14.880	

Source IREF processing on Patronato ACLI data

Table 6 Unemployment risk to the Age of first occupation

Covariates	Haz. Ratio	Std.Err	z	P > lzl	[95% Conf. Interval]
14–18 years	17.79277	1.268047	40.39	0,000	15.47322 20.46005
19–29 years	4.626653	0.3241474	21.86	0,000	4.033025 5.307657
30–39 years	1.143691	0.0811351	1.89	0,058	0.9952294 1.314299
40–49 years	0.7176236	0.0519428	−4.58	0,000	0.6227092 0.827005
50–59 years	0.4563883	0.0330068	−10.85	0,000	0.396072 0.52589
60–69 years	0.527123	0.0388852	−8.68	0,000	0.4561628 0.6091217
Gender (rif. female)	1.120153	0.0088857	14.30	0,000	1.102872 1.137705

Source IREF processing on Patronato ACLI data

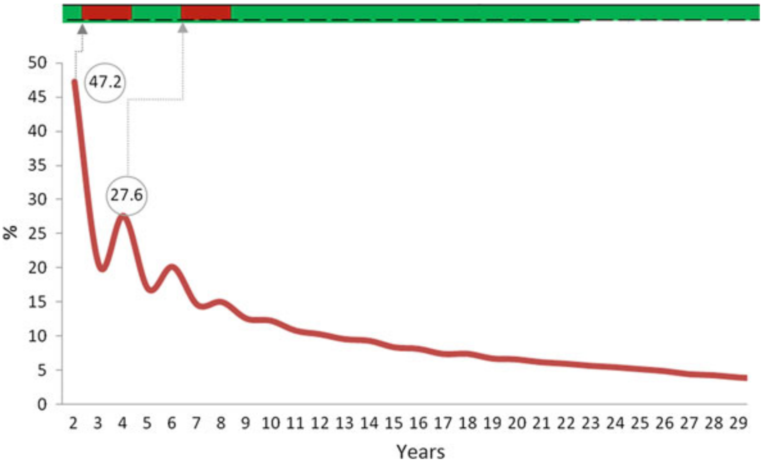


Chart. 1 Unemployment incidence on the transition Employee-Unemployed

"censured cases," workers who lost their jobs during the period of occupation. Basically, the survival analysis in this context of study is used to determine the risk of unemployment over time (Lancaster 1990; Kalbfleisch and Prentice 2002).[7]

Compared to these two groups, it was made a Cox Regression (Cox 1972) selecting as covariates gender and age range of the taxpayer at the time of observation t_0. The Table 6 shows clearly the relationship between the unemployment risk and the age of workers: the hazard ratio is high in the age group 14–18 years (17.8) and 19–29 years (4.6) and decreases with increasing of age of workers. As for gender, men have a risk level of unemployment higher than women (Table 6).

[7] The regression was processed with STATA software on 93,239 social security account statements. The cases censured totaled 64,686. The likelihood parameters are: Log likelihood = -682483.19, LR chi2 = 54571.53 with Prob > chi2 = 0.0000.

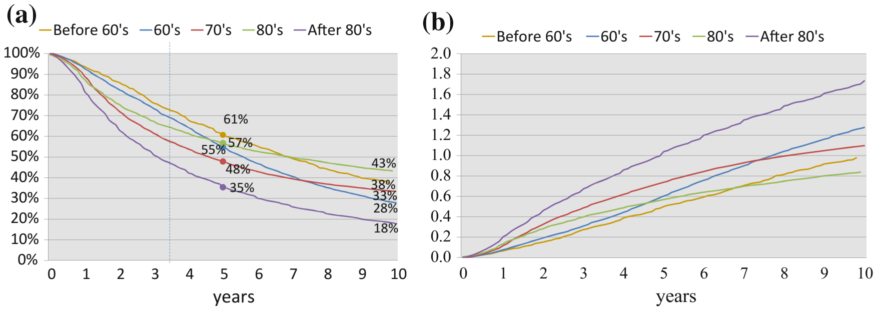


Chart. 2 Survival **a** and unemployment risk (**b**), for pension contribution period

Next to that, it was processed the survival function and risk, using the Kaplan–Meier estimator (Kaplan and Meier 1958; Ramlau-Hansen 1983) for periods of contribution. The survival curves (see Chart. 2a) and unemployment risk (see Chart. 2b) show, in a complementary way, the level of occupational mobility in the various pension contribution periods. At to observing the likelihood of survival is 100%, all subjects are occupied, and the corresponding risk of inactivity is null. Over the years, the percentage of those who retains the first occupation decreases and consequently the risk of inactivity increases, until reaching the maximum value at the end of the observation period (at time t_{10}).

In the comparison between the periods of contribution, the survival analysis shows the high level of occupational mobility of those who began working in the past twenty years: five years after the labor market entry, for these workers the employment level is 35% (see Chart. 2) and falls to 18% after 10 years. It goes without saying that the risk of unemployment for employed people who entered in the last twenty years in the labor market is higher than workers of other contribution periods (see Chart. 2). An important role in the instability of employment of young workers was the introduction, in the nineties, of contractual forms that produced precarious employment: fixed-term contracts, low wages, and reduction of the workers' rights and social protections.

Conversely in the years prior to 1960 (1945–1959) and in the eighties, there was a higher probability of maintaining the first job: before 1960 the employment level within 5 years from the first day of work was 61%, after the eighties was 57%; these values decrease, respectively, to 38% and to 43% after ten years of work. Moreover, if you take into account the difference between the values of the first and the second five years, in the sixties we saw the biggest decline in the number of employees (−29%), a figure that is confirmed by the corresponding risk curve, where we notice a significant increase in the slope of the curve after the fifth year of observation.

3 Conclusions

The brief description of the work carried out on employee benefit statements of the ACLI Patronage highlights the informational potential of the longitudinal analysis. In fact, the longitudinal analysis has highlighted some basic trends that have characterized the labor market in Italy: the centrality of a job model based on employees, the relevant occupational mobility of young people employed beyond the historical reference period and the worker casualization implemented in recent years resulting in an increased risk of inactivity. However, wanting to escape from a focused reading on the results, this working paper is useful for reflecting especially on the state of the art of statistical information in Italy, both in terms of data quality and from the one of their spread and usability.

In terms of quality of statistic information, data analysis of the social security account statements has put in light critical issues linked both to completeness of information and to data consistency; problems that result by a system of management and storage of data "on large meshes," with a low-intensity system control of the data quality. A predominantly bureaucratic-administrative use of data influences the quality of information, and it limits the statistical analysis to a purely descriptive and contingent use. This situation makes more difficult the development of longitudinal models for the study of social phenomena and, at the same time, borders the statistical knowledge linked to the temporality of the objects to a limited audience of insiders.

References

Cleves, M. A., Gould, W. W., Gutierrez, R. G., & Marchenko, Y. V. (2010). *An introduction to survival analysis using stata* (3rd ed.). College Station, TX: Stata Press.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society: Series B, 34,* 187–220.

Fleming, T. R., & Harrington, D. P. (1991). *Counting processes and survival analysis*. New York: Wiley.

Hosmer, D. W., & Lemeshow, S. (2008). *Applied survival analysis: Regression modeling of time to event data* (2nd ed.). New York (NY): Wiley, Spinger.

INPS. (2015). *Non per cassa, ma per equità*. Dicembre: Roma.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). New York: Wiley.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of American Statistical Association, 53,* 457–481.

Lancaster, T. (1990). *The econometric analysis of transition data*. Cambridge: University Press.

Leombruni, R., Quaranta, R., & Villosio, C. (2010). Note di pubblicazione di WHIP v. 3.2, WHIP Technical Report no. 1/2010, LABOR Laboratorio Riccardo Revelli, Centre for Employment Studies, Torino.

Leombruni, R., Richiardi, M., & Costa, G. (2008). Aspettative di vita, lavori usuranti e equità del sistema previdenziale. Prime evidenze dal Work Histories Italian Panel, Laboratorio Riccardo Revelli, Working Paper no. 75. marzo.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics,* 453–466.

Integrated Education Microdata to Support Statistics Production

Maria Carla Runci, Grazia Di Bella and Francesca Cuppone

Abstract To optimize statistics production, National Statistical Offices are using as much as possible administrative data. Currently, in Istat, a project aims to create an infrastructure of integrated administrative microdata on education. For each level of education, available microdata on enrolments and qualifications achievement are integrated in a cross section and in a longitudinal perspective, and qualifications are suitably classified according to national and international standards. The result is a database on education and qualifications, named BIT, that since 2011 is yearly updated using more than 10 datasets per year. Some preliminary quality measurements produced good results; the purpose is to systematize the evaluation process and to define the actions needed to improve the input and the output quality. The BIT may be used for several statistical purposes: to update the population's educational attainment starting from the data collected in the Population Census 2011, to study the educational carriers, to support social survey.

1 New Challenges in Education Statistics

The use of education longitudinal microdata opens up new perspectives for the production of statistics. In the National Statistical Offices, next to the data produced from surveys, exploitation of data deriving from administrative sources is modifying the official statistics production system. The centralized repository of the large amount of integrated administrative microdata acquired by Istat is the SIM (Integrated System of Microdata). This system manages in a standardized way the integration of administrative data that are regularly acquired from various sources

M.C. Runci (✉) · G. Di Bella · F. Cuppone
Istat, Rome, Italy
e-mail: runci@istat.it

G. Di Bella
e-mail: dibella@istat.it

F. Cuppone
e-mail: cuppone@istat.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_26

(tax data, social security data, education data, etc.). Integration takes place through a system of record linkage procedures among units (individuals, economic units, and places) that enter gradually into the system with the periodic administrative data supplies and all the units who are already in it. Procedures end with the assignment of an anonymous identification code to each unit. This code is unique across datasets and across time. The system allows to produced linked data in an efficient way and also satisfies regulatory constraints on confidentiality.

With the aim to maximize the use of administrative data in the perspective of integrated registers, starting from the SIM it is possible to define Activity Registers (Wallgren and Wallgren 2007). In addition to the Employment Register, which already produces statistics in Istat (2014), the Education Register is being developed.

The statistical information needs in terms of education linked microdata are mainly:

- the population's educational attainment;
- the attendance of school and university courses;
- the transition from Education to Work.

Then, the information to be integrated are those relating to: (a) the attendance of a course; (b) the subsequent qualification achievement at the end of each school cycle and academic course. Integration refers to cross-sectional data and longitudinal data among annual administrative data supplies. Identification codes are assigned to individuals, as students enrolled in a course and students that obtain a qualification, and to economic units, as schools and universities that manage the courses. In addition, students can be connected with the future Population Register, and schools and universities are integrated in the Business Register.

Considering longitudinal information on individuals, the transitions from education to work involve both the Education Register and the Employment Register.

2 Integrating Education Microdata to Support Statistics Production

There is no single administrative source covering all kinds of needs, and very accurate analysis of the available administrative sources has been carried out in order to maximize the necessary information coverage.

Currently, 12 different administrative datasets on education have been defined in agreement with the administrative data holder and are received each year by Istat. The main sources are the National Register of School Students, the National Register of University Students, and the PhDs database, all managed by the Ministry for Education, University and Research that provides, with a scheduled timing, data on the courses attendance and the achievement of qualifications.[1] Additional information for special groups of students have to be collected from local administrative sources.

To these datasets, all the tables of the associated metadata must also be added. Starting from this huge amount of information loaded and integrated in the SIM, relevant data are organized in a longitudinal perspective composing a specific database on Education and Qualifications, named BIT, in Italian *Base integrata su Istruzione e Titoli di studio*. From 2011, for the part of the population who is studying at school or university, the BIT aims to describe the educational career and the achievement of the corresponding qualification.

The first version of the BIT used available sources for the years 2010–2012. Soon as the data are available and checked, procedures update the base. Now, it is updated to 2013.

The BIT is composed mainly by two tables: the first one is partitioned per year and includes individual anonymized data on the course of study followed and the educational attainment in the reference year; details on the qualification acquired are collected in the second table. In particular, it contains: attainment date and type of qualification in accord with the official classification system. Finally, the identification codes of school or university where qualification has been acquired are also included allowing to connect data with the schools and universities data, which are part of the more general Economic Units Population in SIM. In this way, information about denomination, type, and localization of the educational institute can be simply derived.

Yearly, about 11 million records are added and integrated in the BIT. The first table of the BIT, partitioned per year, has a total of about 40 million records. The second table of all qualifications acquired since 2011 has about 7 million records, about 4,5 million of diplomas and university degrees coded in detail.

The BIT process allows to update each year the educational attainment of about 2 million of individuals included in the Italian population, starting from the data collected in the Population Census 2011.

The possibility to derive the population's educational attainment variable from administrative data is the main purpose of the BIT. In general, this task is part of the Istat project aimed to maximize the use of administrative data for the next population censuses.

3 Data Classification

To achieve the aim of the BIT, administrative data are classified using the new educational attainments classification adopted for the future Italian Permanent Population Census and also used in the Istat current surveys (e.g., the Labour Force Survey).

[1] Authors thank the Statistical Division of the Ministry for Education, University and Research for the constant and fruitful cooperation.

Table 1 Population census 2011 classification of qualifications: level 1 descriptions and number of categories of more detailed levels

Level 1 description	Level 2 Groups	Level 3 Subgroups	Level 4 Max detail
Diploma of upper secondary education (4–5 years)	6	24	25
Post graduate non-university diploma and AFAM diploma	3	16	16
University degree (2–3 years) old program	15	112	269
Bachelor's degree (first level—new program)	16	180	1.767
Master's degree (second level—old and new program)	14	53	2.350
Total	61	397	4.442

Appropriately ordered, it responds to the need of having an ordinal categorical variable through which it is possible to assign to each individual of the BIT his educational attainment at a given time. This classification is consistent with Isced-A 2011 classification (first digit).

To give continuity to the last Population Census 2011 data dissemination, BIT qualifications are also classified using the very detailed Italian Population Census 2011 coding system, imported, as much as possible, in the BIT in all its levels of detail. It provides hierarchical subdivisions for diplomas and degrees. This second classification task has required the main computational effort, in particular concerning the classification of university qualifications. As it can be seen from Table 1, categories are very detailed (Master's degree has 2.350 categories in Level 4).

Therefore, a specific automated procedure has been implemented for the BIT university qualifications included in the 2011 and 2012 administrative datasets populating the first version of the BIT. The following annual updates are realized using the correspondence table obtained in this phase.

The results of the procedure are satisfactory in terms of efficiency and completeness. Here, they are briefly described.

Census data are linkable with administrative data by a key composed by three variables: type of degree, degree class, and description.

In the 2011 and 2012 administrative data, the 94% of university qualifications had sufficient variables for the linkage procedure and the 99% of these qualifications have been coded in the Level 4 of the Census classification (the most detailed one).

The matching procedure,[2] which has processed 12.135 different keys, is based on the Levenshtein similarity between two strings. In the 79% of cases, this measure was equal to 1—the two descriptions are identical—so the match was made only by the automated procedure; in the remaining cases, the match was defined by a clerical review supported by an automated procedure, comparing strings with the highest similarity value.

[2] The procedure has been implemented by Stefano Daddi.

4 Data Quality Issues

Quality is considered in relation with the Official statistics production process using administrative data (UNECE 2013).

The input evaluation step regards the statistical quality (in terms of statistical usability) of administrative datasets used to build the BIT. The output quality evaluation is the final quality of the BIT, including the integration and the treatment process. It has to be outlined that the BIT is an intermediate product aimed at all internal users as input for their production processes. The input quality framework considered is that derived from Blue Ets international project (Daas and Ossen 2011; Daas et al. 2011; Cerroni et al. 2014). For the output, some of the Eurostat quality dimensions are considered (Eurostat 2015).

Before being integrated into the BIT, each administrative dataset is evaluated, in the beginning to decide the statistical usability and then to monitor the quality. For the input quality evaluation, a Quality Report Card for Administrative data (QRCA) has been defined considering the following quality hyperdimensions and dimensions.

Source Hyperdimension
Administrative dataset identification; Relevance; Privacy and security; Arrangements for the data delivery; Relationships and feedback with administrative source holder
Metadata Hyperdimension
Clarity/Interpretability; Concepts comparability; Concepts temporal stability; Data treatment (by data source holder)
Data Hyperdimension
Technical checks; Integrability; Accuracy; Completeness; Time-related dimension

For each of them, quality indicators and measurement methods are considered. Generally, results show a good overall quality in the three quality hyperdimensions and some limited critical issues that are being treated.

For instance, a good result for the Administrative Pupils Register (school year 2012–13) in the *Data Hyperdimension* comes from the measurement of *Integrability*, considered as the linking variables availability and quality, in terms of amount of missing values: all the 8 variables used for the SIM record linkage process (tax code, name, surname, sex, date of birth, birthplace—composed by municipality, province, country) are available, and most of the records (about the 94%) have no missing values in all items of the 8 variables. In general for the other relevant variables, there is not significant number of missing values in all datasets (*Completeness in terms of variables*).

Some critical issues have been found in the National Register of Pupils regarding the *Completeness in terms of units*: a total undercoverage for the Bolzano province[3] and the Aosta province have been detected. Furthermore, primary school results are missing in 2011, and partially missing in 2012 and 2013.

[3] Bolzano data are acquired separately from 2014 onwards.

Table 2 Percentage of items of the educational attainment variable assigned indirectly, per year

Year	Primary school certificate	Diploma of lower secondary edu.	Diploma of upper secondary edu.	Bachelor's degree	Master's degree
2011	100.00	79.95	0.06	10.86	6.33
2012	72.77	61.91	0.04	1.95	0.01
2013	62.49	46.10	0.05	2.41	0.00

Table 3 Imputed values for the items of the variable school qualifications achieved in 2012

Qualifications	Registered values	Imputed values	Total	Imputed/total %
Primary school	489.022	61.966	550.988	11.2
Lower secondary	545.045	11.327	556.372	2.0
Upper secondary	441.592	3.114	444.706	0.7

Considering data after integration, it has to be said that longitudinal data on the level of education, as ordinal variable, has been very useful both to recover missing or not available data and to check the consistency of data over time.

Table 2 reports the percentage of the educational attainment variable assigned "indirectly" on the base of attendance data, considering the previous level of qualification. In the first year, percentages are higher (for the Primary school all items were imputed because of missing values).

In the following years, these imputations are obviously decreasing as administrative data about qualification achievements are progressively integrated in the BIT. However, some accuracy issues still remain. In the following table, imputed values on school qualifications achieved in the year 2012 with respect to registered data are reported (Table 3).

Comparing longitudinal data on course school attendance referred to 2011–2012 and 2012–2013 school year, where the course level is different, it is possible to derive the qualification achieved at the end of the 2011–2012 school year. The *Imputation rate* is higher for the Primary schools qualifications (11.2%), the Diplomas of Upper secondary school have a very low imputation rate.

In general, for the imputed missing data, there are no details available about school or university issuing the qualifications, data of graduation, etc.

Consistency rules can be derived considering the educational progression; for example, the student career is inconsistent when the qualification acquired in the year T (coming from the School qualifications source) is inconsistent with respect to the enrollment of the school year (T − 1, T) (coming from the School enrollment source). Table 4 reports the number of items which failed at least one editing rule.

The percentage of inconsistent data referred to 2012 is only the 0.1%. To solve inconsistencies, the more direct data of the qualifications achievement were preferred over the enrollment data.

To evaluate the overall BIT qualification *Coverage*, in the Fig. 1, a comparison is reported among available official Ministry macrodata and the BIT distribution of

Table 4 School attendance consistency, year 2012

School attendance consistency	n	%
Consistent	7.192.060	99.89
Inconsistent	8.221	0.11

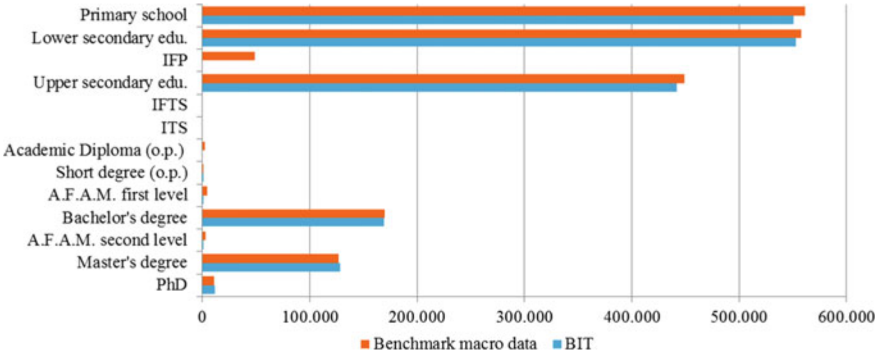


Fig. 1 Qualifications coverage—year 2012

qualifications achieved in 2012 by data level, considering the current classification. The coverage analysis shows a general low undercoverage rate but a total undercoverage for some specific levels.

It is the case of vocational training (IFP) and the higher technical qualifications (IFTS and ITS) issued by Regions and Academic degree of artistic and musical advanced training (AFAM). At this time, for these qualifications, there are no information available at the microdata level from administrative sources.

It should be added that the overall qualifications coverage is going to improve in these last years, for example Bolzano data (School year 2014–15) became available from 2015, Ph.D. data (2012–2014) from 2016. Some actions have been taken to recover the data managed by local authorities.

Finally, about the quality evaluation of the integrating process, the availability of many linking variables of very high quality is a good prerequisite for the success of the procedure. More specific output quality indicators are being studied.

5 Conclusions and Further Developments

The use of educational microdata in a longitudinal perspective opens new scenarios for the production of statistics. The data production process is completely different with respect to the survey data collection. To produce the BIT, education administrative sources have been identified and their actual relevance and usability for statistical purposes have been evaluated. The following step has been the agreement with the administrative source holder about the data acquisition and the practical arrangements in terms of units and variables to acquired, timeliness, data transmission mode. Each administrative dataset supplied is checked in terms of the input quality and integrated into the SIM. Then, a subset of variables enters into the BIT integration process. Qualifications are classified and the output quality is evaluated.

Results are very promising and it seems possible to envisage an increase of efficiency in the education statistics production processes. Considering the high level of detail of the integrated data, the BIT is also an opportunity to produce new statistics on education.

For the next future, the aim is to improve the quality of the BIT. The first goal is to acquire additional data, in particular vocational training qualifications. These qualifications are managed by the Regions, and a comprehensive administrative data source does not exist. The recovery and analysis of the available data will continue hoping to fill this part of undercoverage. The second goal is to increase the degree of data usability for statistical purposes continuing the cooperation with the Ministry for Education, University and Research (reduction of the timeliness, improvement of the checking rules, etc.).

About the output quality measurement, it must be improved and systematized in order to be available for the internal users.

The final aim is to build a completed Education Register able to support the many requirements of the statistical processes: the Population Census increasingly oriented to the use of administrative data; socio-economic surveys which use the educational attainment as a variable of stratification, analysis and dissemination; cohort studies concerning educational careers and transition from education to work.

References

Cerroni, F., Di Bella, G., Galiè, L. (2014). Evaluating administrative data quality as input of the statistical production process. *Rivista di Statistica Ufficiale,* 1-2, 117–146.

Daas, P. J. H., & Ossen, S. J. L. (2011). Metadata quality evaluation of secondary data sources. *International Journal for Quality Research, 5*(2), 57–66.

Daas, P. J. H., Ossen, S., Tennekes, M., Zhang, L. C., Hendriks, C., Foldal Haugen, K., et al. (2011). Reports on methods preferred for the quality indicators of administrative data sources. Deliverable 4.2 of Workpackage 4 of the BLUE-ETS project.

EUROSTAT. (2015). *ESS handbook for quality reports*—2014 edition.

Istat. (2014). Occupazione nelle imprese secondo il nuovo registro Asia-Occupazione, 4 November 2014. Rome: Istat (in Italian).

UNECE. (2013). *The generic statistical business process model generic statistical business process model,* Version 5.0, December 2013.

Wallgren, A., & Wallgren, B. (2007). *Register-based statistics: Administrative data for statistical purposes*. Wiley.

Latent Growth and Statistical Literacy

Emma Zavarrone

Abstract In a world governed by thousands of data, the single number can still create fear among students of behavioral and humanities science. This paradox could be removed through the increase of statistical knowledge at both secondary and post-secondary education levels. Statistical knowledge has been studied in depth in the past, and it is well known as *statistical literacy*. The enhancement of statistical literacy should be realized under new teaching forms more familiar to the digital native generations. This note presents an experiment based on the use of an e-learning tool in order to increase the statistical literacy among Italian students of the humanities. Three cohorts of students were been examined, and the statistical literacy has been tested over time with a dual change difference score model. The results, in limited detail, confirm the initial hypothesis: If the teaching tools are closer to the characteristics of digital native generations, an improvement in statistical literacy can be realized, but the greater the statistical complexity, the less efficient the tools become.

Keywords Statistical literacy · Dual change difference score model · Latent growth · E-learning

1 Introduction

Usually, statistics, within the humanities and behavioral science programs, produces some form of bewilderment among students. Reluctance is based on limited prior knowledge of mathematics which drives to identify the statistics with mathematics. However, the math performance of pupils is not always outstanding such as Pisa–Ocse studies have highlighted in some European regions. So, the automatic thought for the students takes form: "I don't understand mathematics, so I can't study statistics." This thought has a straightforward influence on the final performance in statistics. In 1990, Zeidner (1991) defined *statistics anxiety* and investigated the

E. Zavarrone (✉)
Department Behaviour Marketing and Communication, IULM University, Milan, Italy
e-mail: emma.zavarrone@iulm.it

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-319-55477-8_27

determinants of statistics anxiety: (a) antecedent negative experiences with mathematics,
(b) low performance in mathematics, and (c) a missing sense of math self-efficacy. Similar results were obtained by Onwuegbuzie and Wilson (2003) in which the statistics anxiety was negatively correlated with both the number of statistics course taken and final achievement. A new orientation in teaching statistics Batanero et al. (2011) suggests changes in teaching paradigms, such as the adoption of methods based on real data in order to reduce statistics anxiety. The process is very long, and the anxiety statistics remains. At the same time, the choices of the decision makers are flooded with big data, but an element of doubt becomes increasingly evident. If the population declares that it does not understand statistics, how can they manage a lot of data? Another question follows: Is statistical literacy (hereafter SL) adequate for a basic comprehension of phenomena? Since the 1990s, the international statistical community has dealt with the gap between the high speed of data generation and the low level of SL in the population. Several programs (both national and international) have been promoted in order to facilitate the dissemination of SL especially among young generations. Digital native generations might be the best target for this dissemination because they represent the mix between the classical repertory of study and the digital methods for implementing it. SL engagement among digital natives can be realized through new digital tools such as *apps*. The aim of this note is to test whether the introduction of innovative teaching tools can help students study the statistics, contributing to increase in SL according to the Batanero approach. How does SL grow over time? Sect. 2 is devoted to a literature review on SL, and Sect. 3 focuses on the dual latent change score model, since SL learning is not directly observable. Section 4 presents the data, the results and suggests future directions for research.

2 Statistical Literacy

Statistical literacy has assumed several definitions over time. An former enthusiastic description was been proposed by Ogburn (1940) who disentangled the boundary between maths (a discipline based on calculus) and statistics (a discipline based on measurement) and forecasted a widespread coverage of statistics over the coming years. This prediction was based on the fact that 60% of the US population was attending school and the illiteracy rate for primary school was very low. However, the Ogburn previsions were not widely applied. Walker (1951), inspired by a famous quotation of Wells, introduced the construct of *statistical thinking*, and this construct groups the measures able to interpret the data. During the 1960s, SL was cited many times in the presidential address of the American Statistical Association, but few actions for supporting it have been implemented. In the 1970s and 1980s, the attention on SL was reduced in favor of the statistical thinking approach. Since the 1990s, the statistical thinking has become a component of SL as affirmed by Wallman (1993): "SL is the ability to understand and critically evaluate statistical results that permeate our daily lives coupled with the ability to appreciate the contribution that statistical thinking can make in public and private, professional and personal decisions." The construct of SL went through a transition: from individual informative need to collective need that Gal (2002) focused on "new meaning defining as key ability expected of citizens in information-laden society. Often touted as an expected outcome of schooling and as a necessary component of adults numeracy." The meaning of numeracy tends to overlap with mathematical meaning and refers to SL in the classification elaborated by Delmas et al. (2007) where three dimensions are outlined for managing statistics: *statistical literacy*,which is based on the skills required to understand the basic rules. The measurement is very simple, and it considers the performance of each student; *statistical reasoning* is based on the ability to interpret and critically evaluate statistical information, data-related arguments, and stochastic phenomena, Rumsey (2002); and *statistical thinking* identifies sophisticated abilities used in order to design the studies, choose the methodologies, and interpret the data. Therefore, speaking of statistical literacy means to describe the first level of knowledge in statistics; although a convergence of SL definitions is still missing, the scholars (Batanero et al. 2011; Ben-Zvi and Makar 2015) agree on need to strengthen SL at the secondary school and in university levels.

3 Methodology

The umbrella of latent growth modeling (LGM) collects all models based on longitudinal data in order to quantify the inter-individual variability in intra-individual groups of change over time Curran et al. (2010). The focus on latent growth modeling has ancient roots in Aristotle's thought Zeger et al. (1986). Two centuries ago, Gompertz, Verhult, and Quetelet dealt with the growth population modeling. In the last century, the scholars moved the focus from the growth of observable variables to latent ones Bollen and Curran (2006). The review of the literature was studded with several methodological developments of latent growth modeling, considering Rao (1958) and Tucker (1958) as the first groundbreakers, as they depicted LGM under distinct perspectives (factor and principal component analysis, respectively) and addressed nonlinear modeling, too. Voelkle and Oud (2015), following Grimm and al (2012), proposed to discriminate the LGM based on the nature of time: static versus dynamic time. In the static view, the time has been considered as a predictor, while dynamic time is considered implicitly by the order of the measurement occasions. The dynamic approach refers to the autoregressive cross-lag model Joreskog (1981), latent difference score models McArdle (2001), McArdle and Hamagami (2001), autoregressive latent trajectory models Bollen and Curran (2006), and latent differential models Boker et al. (2004). The dual latent change score model, a special case of latent change score model (LCSM), which was introduced by McArdle (2009), has been applied. LCSM is based on the study of latent differences moving from the well-known test theory in which the observation of *-th* subject at time t (Y_{it}) can be divided into true score (y_{it}) and an error component (e_{it}):

$$Y_{it} = y_{it} + e_{it} \tag{1}$$

The observation at time t – 1 can be written as

$$Y_{it-1} = y_{it-1} + e_{it-1} \tag{2}$$

The difference between y_{it} and y_{it-1} denotes the *latent change score*,

$$\Delta y_t = y_{it} - y_{it-1} \tag{3}$$

The behavior of the latent change score under examination can be reasonably explained over time through an initial true score g_{i0} plus a set of latent changes and a component error, which then becomes:

$$Y_{it} = g_{i0} + \Big(\sum_{t=2}^{T} \Delta y_{it}\Big) + e_{it} \tag{4}$$

The Δy_t can be expressed in three alternative models:

1. *constant*: corresponds to a fixed change over time

$$\Delta y_{it} = g_{i1} \tag{5}$$

2. *proportional*: corresponds to linear model where the changes are directly proportional to the previous latent score, such as

$$\Delta y_t = \beta y_{i,t-1} \tag{6}$$

3. *dual*: corresponds to the composed model of the previous ones where the changes can be divided into two parts: systematic changes (g_{1n}) and proportional change (β) that indicates how each variable influences itself over time.

$$\Delta y_t = g_{it-1} + \beta y_{i,t-1} \tag{7}$$

The dual change difference score model becomes:

$$Y_t = g_{it0} + g_{it-1} + \Big(\sum_{t=2}^{T} \beta y_{i,t-1}\Big) + e_{it} \tag{8}$$

4 Experiment

Due to the lack of interest in quantitative subject, the students enrolled in humanities learn with difficulty the scientific topics and statistics in detail. Currently, in a big data management context, the distance between numbers and words is becoming smaller and smaller. Therefore, the quantitative disciplines have to be treated in a humanities and behavioral science courses, as well. A possible solution could be to substantially reduce this gap between what is perceived as sterile and useless formulas and the everyday context. A conceivable bridge could be represented by a customized system of e-learning, such as MathXL implemented by Pearson Inc., which allows students to complete exercises, provides solutions in real time, and, in the case of errors, receives referrals from the system to the related paragraph of the book. This experiment has been carried out for three academic years (2012/2013, 2013/2014, and 2014/2015) during the Statistics and Market Research (6 ECTS) course in the second year of the Bachelor's degree in Public Relations and Corporate Communication, at the IULM University in Milan, Italy.

4.1 Method and Data

The experiment has been organized by combining two specifics of the platform: the possibility to administer homework (without marks) and a final given test (with marks for correct answers). For every week of lessons, in the platform provides a homework session composed of ten questions and a quiz session always with the same number of tests on the subject covered in class. For the test, the score for every correct exercise was 0.1. The final score ranged from $0(0 \times 0.1)$ and $1(0.1 \times 10)$.

The lease time was constant, even if all the students not equally diligent in terms of the time spent on the assigned tasks.

The left part of Table 1 summarizes the e-learning scores on each time occasion for each cohort. The descriptive statistics for e-learning scores over three academic years denotes a high mean score of the test with very low variability, followed in subsequent times by a slight reduction of the mean score but always with a low variability. These results could be encouraging in terms of understanding the basic statistical concepts. The observed trajectory (Figs. 1, 3, and 5) for each year seems, at least, to describe the same situation: excellent until the middle part of the course (until 4–5 weeks), in which the mean scores are very close to 1 and then worse scores subsequently. Usually, the second part of the course is related to bivariate statistics and to normal distribution, even if from year to year there is a lag between lesson content and the number of weeks. A previous analysis, based on logistic regression, has highlighted a high probability for passing the examination using the e-learning tool, so the research question concerns "how" the students can learn from one period to another? Put differently, given good performance on final examinations, how do students achieve statistical literacy over time? Some assumptions are needed:

Table 1 Descriptive statistics of e-learning scores for time occasions

Times	2012/13	2013/14	2014/15
	N = 309	N = 317	N = 323
	Mean(σ)	Mean(σ)	Mean(σ)
t1	0.95(0.08)	0.94(0.09)	0.98(0.06)
t2	0.97(0.07)	0.97(0.06)	0.96(0.07)
t3	0.94(0.07)	0.97(0.05)	0.96(0.09)
t4	0.96(0.07)	0.96(0.06)	0.97(0.07)
t5	0.94(0.06)	0.98(0.04)	0.96(0.09)
t6	0.97(0.06)	0.97(0.06)	0.89(0.13)
t7	0.90(0.14)	0.92(0.11)	0.94(0.11)
t8	0.87(0.11)	0.96(0.07)	0.83(0.19)
t9	0.77(0.19)	0.91(0.13)	0.76(0.23)
t10	0.71(0.24)	0.86(0.14)	0.94(0.10)
Parameters			
βy	1.069	5.271	7.872
g_0	0.978	0.972	0.971
g_1	−1.586	−5.127	−7.651
σy_{y0}	0.002	0.0001	0.002
σy_{ys}	0.003	0.030	0.130
$\sigma y_{y0,ys}$	−0.002	−0.006	−0.017
Fit statistics			
χ^2(df)	1.313(2)	0.774(2)	1.790(2)
RMSEA	0.000	0.000	0.000
SRMR	0.018	0.014	0.027
CFI	1.000	1.000	1.000
TLI	1.000	1.000	1.000

(a) SL was considered as *numeracy*, (b) the numeracy was assimilated to the latent variable, and (c) previous exploratory factor analysis has showed that one factor accounted for 60% of the total variance in the students' scores(^{a}).

4.2 Results

The dual latent change score model has been applied on three time occasions: t_2, t_4, and t_7 for 2013; t_2, t_6, and t_8 for 2014; and t_2, t_5, and t_8 for the last year (Figs. 1, 2, and 5). These lags, as said before, depend on the heterogeneity of students from one year to next, and this causes delays in the teaching schedule. The right part of Table 1 and Figs. 2, 4, and 6 summarizes the LCSM outputs, in terms of parameter

Fig. 1 2013: observed trajectories

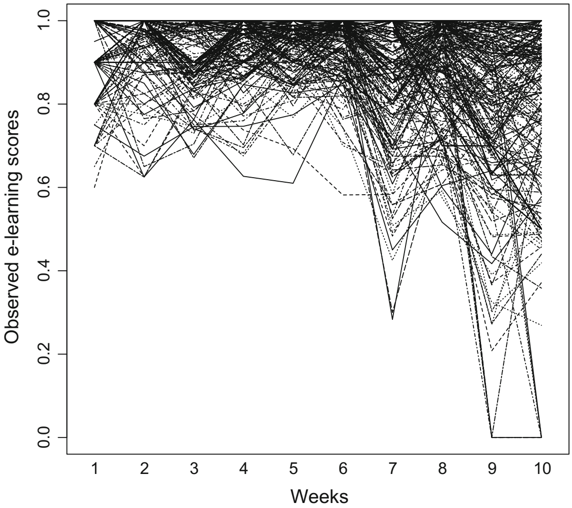
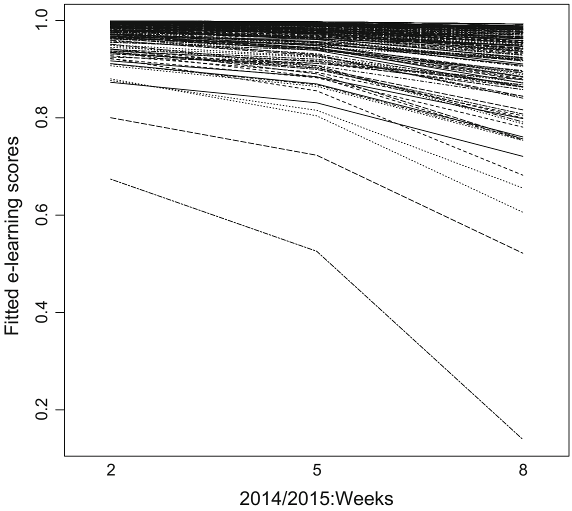


Fig. 2 2013: fitted trajectories



estimate and goodness fit indexes, for each academic year. The findings confirm both the observed trajectory and the form of function that explains the nature of latent change, and all goodness-of-fit indexes denote a good model. On average, the students have received a score from 0.971 to 0.78 for the second quiz. The changes in the numeracy, expressed by βy and g_1, have to be read together, as they denote a negative exponential trend with different rates: For 2013, the first year of experiment, the βy was very small, with only 1% of self-feedback effect on the e-learning scores; in the 2014, the value of βy improves, though the size of constant change was negative, while in 2015, the average e-learning score increased 7% more than

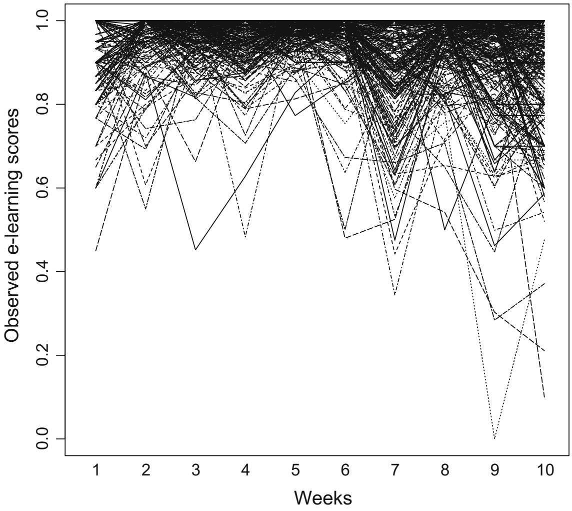


Fig. 3 2014: observed trajectories

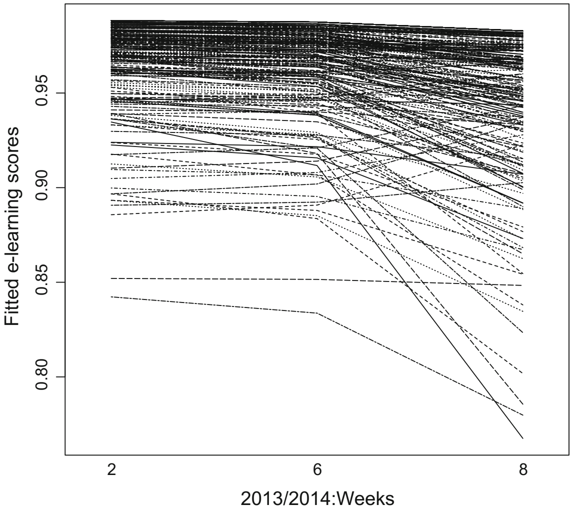


Fig. 4 2014: fitted trajectories

the previous time occasions, but the constant change was negative too. The variance of the latent changes, σy_{ys}, is very low for each of the three years. The covariance between the intercept and constant change component ($\sigma y_{y0,ys}$) is negative and very low, indicating that students with higher true scores at t_2 tended to show no effects in the second part of numeracy learning.

The results show how new tools are requested in order to promote the sequential development of arguments and to manage with these instruments. Further studies will be devoted to better calibrate the adopted techniques.

Fig. 5 2015: observed trajectories

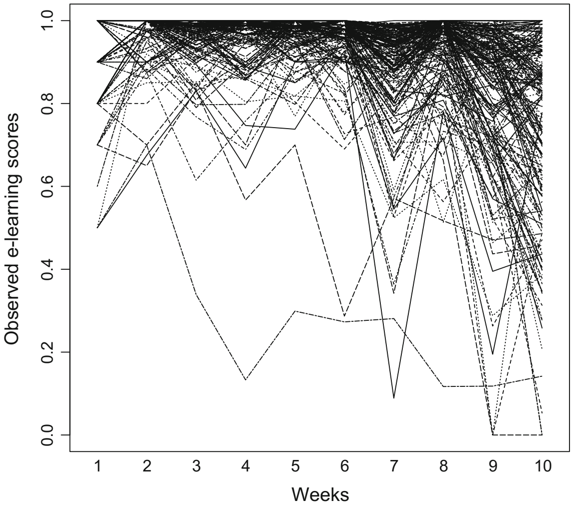
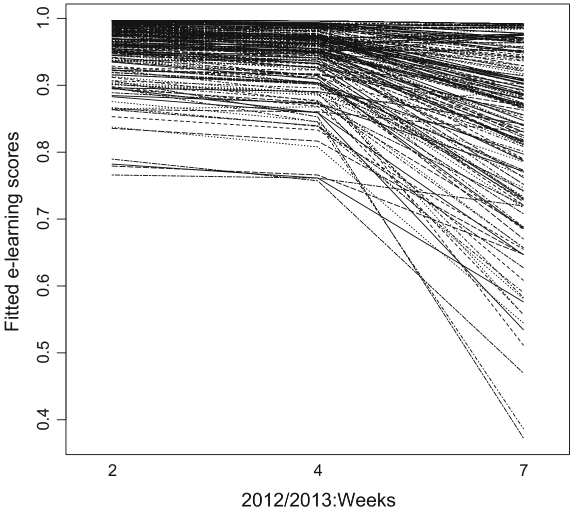


Fig. 6 2015: fitted trajectories



[a] *All results not reported and the R scripts (used packages: lavaan, RAMpath, psych, lattice) are available on request.*

References

Batanero, C., Burrill, G., & Reading, C. (2011). *Teaching statistics in school mathematics*. Challenges for teaching and teacher education: Springer.

Ben-Zvi, D., & Makar, K. (Eds.). (2015). *The teaching and learning of statistics: International perspectives*. Springer.

Boker, S., Neale, M., & Rausch, J. (2004). Latent differential equation modeling with multivariate multi occasion indicators. *Recent developments on structural equation models* (pp. 151–174). Netherlands: Springer.

Bollen, K. A., & Curran P. J. (2006). Latent curve models. In *A structural equation perspective* (Vol. 467). Wiley.

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development*, *11*, 121–136.

Delmas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*, 28–58.

Gal, I. (2002). Adult statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, *70*, 1–25.

Grimm, K. J., et al. (2012). Recent changes leading to subsequent changes. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*, 268–292.

Joreskog, K. G. (1981). Statistical models for longitudinal studies. *Longitudinal Research* (pp. 118–124). Netherlands: Springer.

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analyses. In R. Cudek, S. du Toit, & D. Sorbom (Eds.), *Structural equation modeling: Present and future* (pp. 342–380). Lincolnwood, IL: Scientific Software International.

McArdle, J. J., & Hamagami, F. (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data. In: L. M. Collins, & A. G. Sayer (Eds.). *New methods for the analysis of change. Decade of behavior* (pp. xxiv, 442). Washington, DC, US: American Psychological Association.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*, 577–605.

Ogburn, W. F. (1940). Statistical trends. *Journal of the American Statistical Association*, *35*(209b), 252–260.

Onwuegbuzie, A., & Wilson, V. A. (2003). Statistics anxiety: Nature, etiology, antecedents, effects, and treatments, a comprehensive review of the literature. *Teaching in Higher Education*, *8*(2), 195–209.

Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics*, *14*, 1–17.

Rumsey, D. J. (2002). Statistical literacy as a goal for introductory statistics courses. *Journal Statistics Education*, *10*, 6–13.

Tucker, L. R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika*, *23*, 19–23.

Voelkle, M. C., & Oud, J. H. L. (2015). Relating latent change score and continuous time models. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(3), 366–381.

Walker, H. M. (1951). Statistical literacy in the social sciences. *The American Statistician*, *5*(1), 6–12.

Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association*, *88*, 1–8.

Zeger, S. L., & Harlow, S. D. (1986). Mathematical models from laws of growth to tools for biologic analysis: Fifty years of growth. *Growth*, *51*, 1–21.

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students: Some interesting parallels. *British Journal of Educational Psychology*, *61*, 319–328.

Erratum to: Recurrence Analysis: Method and Applications

Maria Carmela Catone, Paolo Diana and Marisa Faggini

Erratum to:
Chapter "Recurrence Analysis: Method and Applications" in: N. Carlo Lauro et al. (eds.), *Data Science and Social Research*, Studies in Classification, Data Analysis, and Knowledge Organization, https://doi.org/10.1007/978-3-319-55477-8_14

The original version of the book was inadvertently published without the author name "Paolo Diana" in the chapter "Recurrence Analysis: Method and Applications". The author's name has now been included in the chapter and the book has been updated with the change.

The updated online version of this chapter can be found at
https://doi.org/10.1007/978-3-319-55477-8_14

N. Carlo Lauro et al. (eds.), *Data Science and Social Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-319-55477-8_28