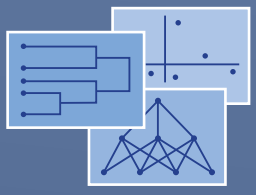
Studies in Classification, Data Analysis, and Knowledge Organization

Paolo Mariani · Mariangela Zenga *Editors*

# Data Science and Social Research II

## Methods, Technologies and Applications



Springer

# Studies in Classification, Data Analysis, and Knowledge Organization

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

Paolo Mariani · Mariangela Zenga
Editors

# Data Science and Social Research II

## Methods, Technologies and Applications

Springer

*Editors*
Paolo Mariani
Department of Economics
Management and Statistics
University of Milano-Bicocca
Milan, Italy

Mariangela Zenga
Department of Statistics
and Quantitative Methods
University of Milano-Bicocca
Milan, Italy

ISSN 1431-8814 ISSN 2198-3321 (electronic)
Studies in Classification, Data Analysis, and Knowledge Organization
ISBN 978-3-030-51221-7 ISBN 978-3-030-51222-4 (eBook)
https://doi.org/10.1007/978-3-030-51222-4



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

# Preface

As digital technologies, the Internet and social media become increasingly integrated into society, a proliferation of digital footprint of human and societal behaviours are generated in our daily lives. All these data provide opportunities to study complex social systems, by the empirical observation of patterns in large-scale data, quantitative modelling and experiments. The social data revolution has not only produced new business models, but has also provided policymakers with better instruments to support their decisions.

This book consists of a selection of the papers presented at the Second International Conference on Data Science and Social Research held in Milan in February 2019 (https://www.dssr2019.unimib.it). The conference aimed to stimulate the debate between scholars of different disciplines about the so-called data revolution in social research. Statisticians, computer scientists and experts on social research discussed the opportunities and challenges of the social data revolution to create a fertile ground for addressing new research problems.

The volume collects 30 contributions focused on the topics for complex social systems. Several papers deal in new methodological developments to extract social knowledge from large scale data sets and new social research about human behaviour and society with large datasets, either mined from various sources (e.g. social media, communication systems) or created via controlled experiments. Moreover, some contributions analysed integrated systems to take advantage of new social data sources; others discussed big data quality issues, both as a reformulation of traditional representativeness and validity and as emerging quality aspects such as access constraints, which may produce inequalities.

All contributions were subjected to peer-review and are listed in alphabetical order of the first author.

We would like to express our gratitude to the Scientific, Program and Local Committees that allowed the realisation of the conference. Moreover, we would thank the authors and the anonymous referees who made the creation of the volume possible. Our deep gratitude also goes to Laura Benedan that supported us in every organisational stage of this book.

Milan, Italy
February 2020

Paolo Mariani
Mariangela Zenga

# Contents

# Digital Methods and the Evolution of the Epistemology of Social Sciences


**Enrica Amaturo and Biagio Aragona**

**Abstract** After ten years that the debate on big data, computation and digital methods has been a contested epistemological terrain between some who were generally optimistic, and others who were generally critical, a large group of scholars, nowadays, supports an active commitment by social scientists to face the digital dimension of social inquiry. The progressive use of digital methods needs to be sustained by an abductive, intersubjective and plural epistemological framework that allows to profitably include big data and computation within the different paradigmatic traditions that coexist in our disciplines. In order to affirm this digital epistemology it is critical to adopt a methodological posture able to elaborate research designs with and against the digital, trying to exploit what digital techniques can give as added value, but going to test their reliability, alongside others techniques, including qualitative ones.

## 1 Introduction

Big data, computation, and digital methods have been a contested epistemological terrain for the last decade of social research. In this controversy, oversimplifying, we have had two groups, with different postures on the matter. On one side, those who were generally optimistic, on the other side, those who were generally critical (Salganik 2018).

Since the advent of big data, the epistemological debate has then developed between two opposites: revolution and involution. According to revolution, disruptive technical changes transform in better both the sciences and the methods once consolidated within the different scientific disciplines (Lazer 2009; Mayer-Schonberger and Cuckier 2012). On the contrary, involution states that digital methods and big data impoverish social sciences and their method (Boyd and Crawford 2012), becoming

E. Amaturo · B. Aragona (✉)
University Federico II of Naples, Naples, Italy
e-mail: aragona@unina.it



P. Mariani and M. Zenga (eds.), *Data Science and Social Research II*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-030-51222-4_1

a threat to the empirical sociology based on surveys and interviewing (Savage and Burrows 2007).

These views developed around two main topics: the quality of digital data, and the role of technology in social research. "Revolutionary" scholars believe that more data automatically lead to better research. This actually is not sustained by facts. Loads of data may actually increase the level of "noise" in data, so we can hardly distinguish rumors form signals (Torabi Asr and Taboada 2019). Moreover, more data does not mean better data *tout court*, because the risk of *garbage in - garbage out* is higher (Kitchin and Lauriault 2015). For these reasons, scholars who maintain involution are critical with the quality of digital data sources, which are mainly secondary data repurposed with new objectives. In analog research most of the data that were used for social research were created with the purpose of research, while in digital social research huge amount of data are created by corporations and organizations with the intent of making profit, delivering services, or administering public affairs.

In a similar way, optimistics are enthusiast about the possibilities opened by digital technology, while critics contrast the adoption of technology. Scholars who sustain the revolution in social sciences believe that technology is the driver of innovation and of advances in knowledge. This technological determinism promotes an idea in which the scientific disciplines stands at the passive end (Marres 2017), with technology being a force of improvement for research. On the contrary, most critics believe that the reconfigurability of digital infrastructures and devices is contested and it needs continue demonstration and testing. Big data and digital methods are effective as far as we would be in the condition to inspect the theoretical assumptions within the data and the socio-technical processes that shape them.

After ten years that the debate on big data and digital methods has developed on these two opposite visions, a large group of scholars, nowadays, supports an active commitment by social scientists to face the digital dimension of social inquiry (Orton-Jhonson and Prior 2013; Lupton 2014; Daniels and Gregory 2016). On this basis, instead of opposing revolution and involution, it should be more correct to talk about an epistemic evolution. These scholars advocate that in order to effectively improve the unfolding of social phenomena, digital methods and big data have to be integrated with traditional data sources and methods already existing in social sciences. The works developed by these scholars focus on the definition of an epistemology of social research that adopts a critical posture on the role that digital technology must have in scientific research, but, at the same time, creative on the possibilities offered by technology to research (Marres 2017; Halford and Savage 2017).

In this article, we argue that one way to do that is concentrating more efforts on the construction of research designs for and against the digital. With the objective to avoid ideological positions about the role of digital methods in social research, we should then promote an active engagement in testing the different instruments of digital research, such as big data, machine learning, platforms analytics, search as research tools, and so on. Moreover, digital technologies carry the risk of flattening social research only on two phases: data analysis, and communication of results. The development of techniques for elaborating increasingly large databases—which often are not even fully understood by social scientists (Kitchin 2014a)—pushes to

consider the process of research as an analytical process, effectively neglecting the other research phases. Likewise, data visualization and infographics, which are so fundamental to digital data communication, can produce the same risks (Halford and Savage 2017). One possible answer to these concerns is again to invest more efforts in developing digital research designs. By refocusing the attention on the research design we can restore importance to all the other research phases.

## 2 Revolution

The idea that digital methods, computation and big data are revolutionary for social sciences has developed upon three main epistemic features: objectivity, induction and computation.

The objectivity leaves from the fact that reality is considered independent from technology and sociality. Big data, platforms and digital ecosystems are seen as windows on the social reality. The scientific method is data driven, and—resting on the possibility to track human behavior with unprecedented fidelity and precision—exploring existing data may be more useful than building models on why the people behave the way they do. The objectivity of reality has the further consequence of repurposing the dualism between the researcher and the reality, between subject and object, that is typical of positivism. What is different with early positivism is the objective of research, which is more centered on quantification and description than on looking for causes of phenomena. Data are the core of this approach to social science, and computational methods are the required tools to learn from these data. This view on the digital is often espoused by scholars in quantitative computational social sciences (Lazer 2009) and data science.

The revolutionary idea has been also sustained by discussions about innovation in social research, which are more methodological than epistemological (i.e. if digital techniques should be considered new or not (Marres 2017). The new-vs-old dichotomy was firstly promoted by Rogers (2013), which distinguished between digitized and digital native techniques. The former are those that already existed in analog form, and that are "migrated" on the web (for example web surveys and digital ethnography), the latter are those "born" on the web, such as web scraping techniques, and search as research tools. The division of digital methods in these two groups has at least two weaknesses that may be envisaged. First of all, those who advocate the existence of digital native techniques propose the idea of methodological development as guided by technology. Technology = new; social sciences = old. Furthermore, this opposition implies that the development of research methods in social sciences should come from the "outside", from disciplines such as computer science and data science. But, all the techniques that are having great relevance in digital research: "have an inter-disciplinary origin … and can be qualified as" mixed "techniques, in the sense that they combine computational elements and sociological elements" (Marres 2017, p. 104). Halford (2013) believes that digital techniques are not at all "alien" to social sciences, but rather that the techniques incorporated

in digital platforms and devices are built on consolidated and lasting methodological principles. By stressing the revolution of social sciences and their methods, scholars focus on ruptures instead than on connections, with the result that the continuity between traditions of social research and digital techniques will no longer be recognized.

In order to clarify this point, there is the example of the analytics that have been developed by *Google* to do research through searches made on its search engine. By using these analytics, such as *Google Trends*, Google carried out a famous study on the ability to predict the propagation of influenza before the research institute (CDC) that was responsible for measuring its spread. That work (Ginsberg et al. 2009) was used to state that digital native techniques enabled forms of analysis that could not be realized before (Mayer-Schonberger and Cuckier 2012; Rogers 2013). Abbott (2011), however, noted that these tools rested on very traditional forms of analysis. For example, *Trends* uses the analysis of temporal and territorial series to count how many times keywords have been searched on the search engine. That is, even if the technique is innovative with respect to the technological and computational aspects, the underlying methodological principle is very old. It would then make more sense, even at the methodological level, to appropriately examine how digital techniques lead/do not lead to social research method evolution, rather than focusing the attention on its revolution.

## 3 Evolution

It is not difficult to support the idea that big data, digital methods and computation contribute to an evolution of social research methods, and more generally of social sciences. The spreading of large databases from a variety of sources gives the possibility of doing research in many ways, and of improving techniques that were already used in the past (i.e., content analysis and network analysis). At the same time, critical data science research (Iliadis and Russo 2016) is emerging, with the aim to assess the social consequences of the processes of digitalization, and consequent datafication, of various sectors of society. The collection, analysis and processing of data, networks and relationships through digital methods therefore manages to create also new points of contact between digital and not digital social science (Orton-Jhonson and Prior 2013; Daniels and Gregory 2016).

This evolutionary posture starts with an active commitment by social scientists to confront with the technological dimension of social inquiry. Its main features are: intersubjectivity, abduction and mixed methods.

Intersubjectivity refers to the fact that social reality is dependent on the socio-technical activities that are made to grasp it. Data are not a simple reflection of a world that "is", but are thoroughly "produced". The separation between object and subject must be overcome. It acknowledges the role of platforms (Van Dijck et al. 2018) and «methodological dispositifs» (Ruppert 2013) in shaping reality.

Digital platforms are changing with a velocity that is not usual for social science data (Chandler et al. 2019). For example, with longitudinal survey data, breaks in the series are rare and very carefully implemented inside the overall longitudinal research design. On the contrary, platforms change all the time, and changes occur in at least three ways (Salganik 2018): they may change in who is using them, in how they are used, and in how the platform work. Some examples are: during 2012 US Presidential Election the proportions of tweets by women increased and decreased from day to day; *Facebook* in Italy started to be a social network to reconnect the school community, and now is used also as a form of advertising; in 2018 *Twitter* decided to double the number of digits in tweets from 140 to 280. All these kinds of changes may have an impact on research results.

Also the "methodological dispositifs" may impact on results. "Methodological dispositifs" are the material objects and ideas that configure the ways we do research. They are not simply research methods, but they are also the same objects of analysis. To understand the role played by these dispositifs a close inspection of data assemblages should be realized (Kitchin 2014b; Aragona 2017). Data assemblage is a complex socio-technical system composed of many elements and elements that are thoroughly entwined, whose central concern is the production of data. These assemblages are made of two main activities: a technical process, (operational definitions, selection, data curation) which shape the data as it is, and a socio-cultural process, which shapes the background knowledge (beliefs, instruments and other things that are shared in a scientific community). Researching big data assemblages may help to unpack digital black boxes (Pasquale 2015) and increase our knowledge about the processes of algorithms construction (Aragona and Felaco 2018), the effects of data curation on research results (Aragona et al. 2018), the values into data.

Moreover, according to evolution, neo-empiricism—the data-first method—has to be rejected. Social sciences must preserve the main tenets of the post-positivist scientific method, but at the same time promote the joint use of induction and deduction. For the advocates of evolution, scientific knowledge is pursued using "guided" computational techniques to discover hypotheses to be submitted to subsequent empirical control. The process is guided because the existing theories are used to direct the development of the discovery and not—as in quantitative computational science—to identify all the existing relationships in a database. Instead, the way in which data is constructed or re-analyzed is guided by some assumptions, supported by theoretical and practical knowledge and experience of how technologies and their configurations are able to produce valid and relevant empirical material for research. In practice, the method used is abductive (Pierce 1883), and aims to insert unexpected results in an interpretative framework.

Consequently, also the opposition between correlation and causation can be overcome. Despite the fact that quantitative computational social science has become the most widespread way of doing computational social science, it should not be forgotten that the ambitions of the authors who wrote the *Manifesto of Computational Social Science* (Conte et al. 2012) were different. Conte (2016) underlines that at the beginning there was no quantitative approach, but computational social science was mainly generative and aimed to unveil the mechanisms that produce

social phenomena through simulations in informational ecosystems. This way of doing CSS has produced many theories about social phenomena such as cooperation, coordination and social conventions. Quantitative CSS, instead, it dismissed the search for the causes of social phenomena. The theoretical ambitions of the authors of the *Manifesto* have been supplanted by an emphasis on quantification and description, mainly because, as Merton first noted (1968), science goes to sectors where there is abundance of data.

Finally, evolutionary social science epistemology assumes that more or less computational analytical methods have become the standard of social research, but at the same time, it does not consider it an imperative. It may be useful to produce new visions of social phenomena through digital methods, but their methodological capacity should be constantly tested. The use of these techniques must be openly discussed, evaluating the impacts on research designs, on the formulation of questions and, when necessary, on hypothesis testing strategies. The already cited *Google Flu Trends* research is a good example of that. After early enthusiasm, the use of search queries to detect the spread of flu turned to be tricky. Over time, researchers discovered that the estimates were not so much better than that of a simple model that calculates the amount of flu based on a linear extrapolation from the two most recent measurements of flu prevalence (Goel et al. 2010). Moreover, estimates were prone to short-term failure and long-term decay. More specifically, during the Swine Flu pandemic, the trends overestimated the amount of influenza, probably because people change their search behavior during a global pandemic. It was only thank to the control of their results with those that are collected by the US Centers for Disease Control (CDC), which are regularly and systematically collected from carefully sampled doctors around US, that researchers were able to develop more precise estimates. Studies that combine big data sources with researcher-collected data will enable companies and governments to produce more accurate and timely measures.

## 4 Conclusions

If social sciences want to benefit from the opportunities of big data, computation and digital methods, the path to follow is adopting an epistemological perspective that not only overcomes the revolution-involution dichotomy, and the new methods-old methods one, but that also call to question some of the main dichotomy that have characterized epistemology of social sciences since recently, such as: subject-object; induction-deduction; correlation-causation.

First of all, intersubjectivity, and the attention to the context in which the representations of phenomena to be investigated are realized, constitute a fundamental starting points for an epistemology of the digital. Indeed, digital technologies have confirmed that the objects of study, and the subjects who study, both actively co-construct data. As highlighted by Lupton (2014), from the moment in which digital

research techniques are used, they are theorized. Therefore, it is not possible to separate the digital analysis as an object of study, from the analysis with digital techniques itself, because both require focusing on the ways in which they are co-constituted.

A second point, linked to the first, is that a digital epistemology that wants to avoid the simplistic positions of neoempirists must pay more attention to the process. Although big data and computational techniques are able to analyze social phenomena in real time, most of the digital data represents a set of snapshots of events that update very quickly. Nothing that registers big data can capture the processes or mechanisms that determine the changes that are detected by the data (O'Sullivan 2017). Causation cannot be obtained exclusively through big data.

Moreover, technological determinism should be overcome. Digital methods may be an interesting and promising way to inspire social sciences only if we are able to inspect the theoretical premises that are embedded in the data, and the socio-technical processes that determined their final form. Recognizing the role of technology in the configurations of social research does not imply technological determinism, and that technology must guide scientific knowledge.

Epistemology of the digital needs to become concrete through the definition of a creative and critical method, that elaborates research designs with and against the digital (Marres 2017). These designs try to exploit what digital techniques can give as added value, but at the same time are going to test their reliability, alongside others techniques, including qualitative ones.

It is only in an abductive, intersubjective and critical epistemological framework and through a mixed and creative method that the current technological character of the digital social inquiry can be profitably conveyed within the different paradigmatic traditions that coexist in our disciplines.

## References

Abbott, A. (2011). *Google of the past. Do keywords really matter?* Lecture of the Department of Sociology, Goldsmith, 15th March.

Aragona, B. (2017). New data science: The sociological point of view. In E. Amaturo, B. Aragona, M. Grassia, C. Lauro, & M. Marino (Eds.), *Data science and social research: Epistemology, methods, technology and applications.* Heidelberg: Springer.

Aragona, B., & Felaco, C. (2018). The socio-technical construction of algorithms. *The Lab's Quarterly*, *44*(6), 27–42.

Aragona, B., Felaco, C., & Marino, M. (2018). The politics of Big Data assemblages. *Partecipazione e conflitto*, *XI*,(2), 448–471.

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

Chandler, J., Rosenzweig, C., & Moss, A. J. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods*, *51*, 2022–2038.

Conte, R. (2016). Big Data: un'opportunità per le scienze sociali? *Sociologia e Ricerca Sociale*, *CIX*, 18–27.

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., et al. (2012). Manifesto of computational social science. *The European Physical Journal Special Topics CCXIV*, 325–346.

Daniels, J., & Gregory, K. (Eds.). (2016). *Digital sociologies*. Bristol: Policy Press.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature CDLVII, 7232,* 1012.

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.

Halford, S., Pope, C., & Weal, M. (2013). Digital futures? Sociological Challenges and opportunities in the emergent semantic web. *Sociology XLVII, 1,* 173–189. https://doi.org/10.1177/0038038512453798

Halford, S., & Savage, M. (2017). Speaking sociologically with big data: Symphonic social science and the future for big data research. *Sociology*, *LI*(6), 1132–1148.

Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society*, *I*(2), 1–8.

Kitchin, R. (2014a). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *I*(1), 1–12.

Kitchin, R. (2014). *The data revolution: Big Data Open Data Data Infrastructures and Their Consequences*. London: Sage.

Kitchin, R., & Lauriault, T. P. (2015). Small data in the era of big data. *GeoJournal*, *80*(4), 463–476.

Lazer, D., Brewer, D., Christakis, N., Fowler, J., & King, G. (2009). Life in the network: The coming age of computational social science. *Science CCCXXIII, 5915,* 721–723.

Lupton, D. (2014). *Digital sociology*. London: Routledge.

Marres, N. (2017). *Digital sociology: The reinvention of social research*. New York: Wiley.

Mayer-Schönberger, V., & Cukier, K. (2012). *Big Data: A revolution that transforms how we work, live, and think*. Boston: Houghton Mifflin Harcourt.

Merton, R. K. (1968). *Social theory and social structure*. Glencoe (IL): Free Press.

Orton-Johnson, K., & Prior (a cura di), N. (2013). *Digital sociology: Critical perspectives*. Heidelberg: Springer.

O'Sullivan, D. (2017). *Big Data: why (oh why?) this computational social science?* www.escholarship.org. Accessibile all'URL https://escholarship.org/uc/item/0rn5n832. Ultimo accesso 1 febbraio 2018.

Pasquale, F. (2015). *The black box society*. Cambridge (MA): Harvard University Press.

Peirce, C. S., & (a cura di), . (1883). *Studies in logic, Boston (MA)*. Brown and Company: Little.

Rogers, R. (2013), *Digital methods*. Cambridge (MA): MIT press.

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

Salganik, M. J. (2018). *Bit by bit: Social research in the digital age*. London: Princeton.

Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*, *41*(5), 885–899.

Torabi Asr, F., & Taboada, M. (2019). Big Data and quality data for fake news and misinformation detection. *Big Data & Society*, *6*(1).

Van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford: Oxford University Press.

# Restricted Cumulative Correspondence Analysis


**Pietro Amenta, Antonello D'Ambra, and Luigi D'Ambra**

**Abstract** In the context of the non-iterative procedures for performing a correspondence analysis with linear constraints, a new approach is proposed to impose linear constraints in analyzing a contingency table with one ordered set of categories. At the heart of the approach is the partition of the Taguchi's statistic which has been introduced in the literature as simple alternative to Pearson's index for contingency tables with an ordered categorical variable. It considers the cumulative frequency of cells in the contingency tables across the ordered variable. Linear constraints are then included directly in suitable matrices reflecting the most important components, overcoming also the problem of imposing linear constraints based on subjective decisions.

## 1 Introduction

Correspondence Analysis (CA) is a popular tool to obtain a graphical representation of the dependence between the rows and the columns of a contingency table (Benzecri 1980; Greenacre 1984; Lebart et al. 1984; Nishisato 1980; Beh 2004; Beh and Lombardo 2012, 2014). This representation is obtained by assigning scores in the form of coordinates to row and column categories. CA is usually performed by applying a singular value decomposition to the matrix of the Pearson ratios or the

P. Amenta (✉)
Department of Law, Economics, Management and Quantitative Methods,
University of Sannio, Benevento, Italy
e-mail: amenta@unisannio.it

A. D'Ambra
Department of Economics, University of Campania "L.Vanvitelli", Capua , Italy
e-mail: antonello.dambra@unicampania.it

L. D'Ambra
Department of Economics, Management and Institutions ,
University of Naples "Federico II", Naples, Italy
e-mail: dambra@unina.it



P. Mariani and M. Zenga (eds.), *Data Science and Social Research II*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-030-51222-4_2

standardized residuals of a two-way contingency table. This decomposition ensures that the maximum information regarding the association between two categorical variables are accounted for in one or two dimensions of a correspondence plot. However, little attention in literature has been paid to the case where the variables are ordinal. It is well known that the Pearson chi-squared statistic (likewise CA) can perform poorly in studying the association between ordinal categorical variables (Agresti 2007; Barlow et al. 1972). An approach dealing with this theme (Beh et al. 2011), in a CA perspective, is based on the Taguchi's statistic (Taguchi 1966, 1974) considering the cumulative frequency of cells in the contingency tables across the ordered variable. Taguchi's statistic has been introduced in the literature as a simple alternative to Pearson's index for contingency tables with an ordered categorical variable. Beh et al. (2011) developed this variant of CA in order to determine graphically how similar (or not) cumulative (ordinal) response categories are with respect to (nominal) criterion ones.

Note that the interpretation of the multidimensional representation of the row and column categories, for both approaches, may be simplified if additional information about the row and column structure of the table is available and incorporated in the analysis (Böckenholt and Böckenholt 1990; Takane et al. 1991; Böckenholt and Takane 1994; Hwang and Takane 2002). Differences between constrained and unconstrained solutions may highlight unexplained features of the data in the exploratory analyses of a contingency table. In the classical analysis, Böckenholt and Böckenholt (1990) (B&B) considered the effect of concomitant variables (given by the external information) partialling them out from the CA solution according to the null-space method. The aim of this paper is to consider an extension of the B&B's approach (Böckenholt and Böckenholt 1990) to contingency tables with one ordered set of categories by using additional information about the structure and association of the data. This extension is achieved by considering the variant of CA based on the decomposition of the Taguchi's statistic (Beh et al. 2011). A new explorative approach named *Restricted Cumulative Correspondence Analysis* is then introduced.

## 2 Basic Notation

Consider a two-way contingency table $\mathbf{N}$ describing the joint distribution of two categorical variables where the $(i,\ j)$-th cell entry is given by $n_{ij}$ for $i = 1, ..., I$ and $j = 1, ..., J$ with $n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$. Let $n_{i\bullet}$ and $n_{\bullet j}$ be the $i$th row and $j$th column marginal frequencies, respectively. The $(i,\ j)$-th element of the probability matrix $\mathbf{P}$ is defined as $p_{ij} = n_{ij}/n$ so that $\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} = 1$. Suppose that $\mathbf{N}$ has one ordered set of categories with row and column marginal probabilities given by $p_{i.} = \sum_{j=1}^{J} p_{ij}$ and $p_{.j} = \sum_{i=1}^{I} p_{ij}$, respectively. Moreover, let $\mathbf{D}_I$ and $\mathbf{D}_J$ be the diagonal matrices whose elements are the row and column masses $p_{i.}$ and $p_{.j}$, respectively. Lastly, $z_{is}$ is the cumulative frequency of the $i$-th row category up to the $s$-th column category.

## 3 Visualizing the Association Between a Nominal and an Ordinal Categorical Variable

CA of cross-classifications regarding the association (using $X^2$ as its measure) between two categorical variables has been used by the data analysts from a variety of disciplines over the past 50 years. It is a widely used tool to obtain a graphical representation of the dependence between the rows and columns of a contingency table. CA can be usually performed by applying the Singular Value Decomposition (SVD) on the Pearson's ratios table $\tilde{\mathbf{P}} = \mathbf{D}_I^{-1/2}\mathbf{P}\mathbf{D}_J^{-1/2}$ (Goodman 1996) of generic term $\alpha_{ij} = p_{ij}/p_{i.}p_{.j}$ with $i = 1, \ldots I$ and $j = 1, \ldots, J$. That is, for the $I \times J$ correspondence matrix $\mathbf{P}$, CA amounts to the decomposition $\mathbf{D}_I^{-1/2}\mathbf{P}\mathbf{D}_J^{-1/2} = \tilde{\mathbf{A}}\tilde{\Delta}\tilde{\mathbf{B}}^T$ with $\tilde{\mathbf{A}}^T\tilde{\mathbf{A}} = \mathbf{I}$, $\tilde{\mathbf{B}}^T\tilde{\mathbf{B}} = \mathbf{I}$ and $\tilde{\Delta} = \text{diag}(1, \lambda_1, ..., \lambda_K)$ where the singular values $\lambda_m$ are in descending order ($m = 1, \ldots, K$ with $K = \min(I, J) - 1$) and matrices $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ contain the left and the right singular vectors, respectively. If we omit the trivial solutions then CA amounts to the SVD of the matrix

$$\boldsymbol{\Pi} = \mathbf{D}_I^{-1/2}(\mathbf{P} - \mathbf{D}_I\mathbf{1}\mathbf{1}^T\mathbf{D}_J)\mathbf{D}_J^{-1/2} = \mathbf{A}\boldsymbol{\Lambda}\mathbf{B}^T$$

with $\mathbf{A}^T\mathbf{A} = \mathbf{I}$, $\mathbf{B}^T\mathbf{B} = \mathbf{I}$ and $\Lambda$ diagonal matrix where singular values $\lambda_m$ are in descending order. The theoretical developments and applications of CA have grown significantly around the world in nearly all disciplines. However, little attention in literature has been paid to the case where the variables are ordinal. Indeed, Pearson's chi-squared statistic test of independence between the variables of a contingency table does not perform well when the rows/columns of the table are ordered (Agresti 2007, Sect. 2.5; Barlow et al. 1972).

Taguchi's statistic has been introduced in the literature (Taguchi 1974, 1966; Nair 1986, 1987) as simple alternatives to Pearson's index for contingency tables with an ordered categorical variable. Taguchi's statistic takes into account the presence of an ordinal categorical variable by considering the cumulative sum of the cell frequencies across the variable. To assess the association between the nominal and ordered column variables, Taguchi (1966, 1974) proposed the following statistic

$$T = \sum_{s=1}^{J-1} \frac{1}{d_s(1-d_s)} \sum_{i=1}^{I} n_{i\bullet}\left(\frac{z_{is}}{n_{i\bullet}} - d_s\right)^2 \tag{1}$$

with $0 \leq T \leq n(J-1)$ and where $d_s = \sum_{i=1}^{I} z_{is}/n = z_{\bullet s}/n$ is the cumulative column proportion up to $s$-th column. Both Nair (1986) and Takeuchi and Hirotsu (1982) showed that the $T$ statistic is linked to the Pearson chi-squared statistic so that $T = \sum_{s=1}^{J-1} X_s^2$ where $X_s^2$ is Pearson's chi-squared statistic computed on the generic contingency tables $\mathbf{N}_s$ of size $I \times 2$. This table is obtained by aggregating the columns (categories) $1, \ldots, s$ and the remaining ones $s+1, \ldots, J$ of table $\mathbf{N}$, respectively. For this reason, (Nair 1986) referred to Taguchi's statistic $T$ as the

*cumulative chi-squared statistic* (CCS). By generalizing (1), Nair (1986) considers the class of CCS-type tests

$$T_{CCS} = \sum_{s=1}^{J-1} w_s \left[ \sum_{i=1}^{I} n_{i\bullet} \left( \frac{n_{is}}{n_{i\bullet}} - d_s \right)^2 \right] \tag{2}$$

and corresponds to a given set of weights $w_s > 0$. The choice of different weighting schemes defines the members of this class. A possible choice for $w_s$ is to assign constant weights to each term ($w_s = 1/J$), Nair (1986, 1987) shows that, for this choice, the statistic $T_{CCS}$ becomes $T_N = \frac{1}{J} \sum_{s=1}^{J-1} \sum_{i=1}^{I} N_{i\bullet} \left( \frac{Z_{is}}{N_{i\bullet}} - d_s \right)^2$ and has good power against ordered alternatives. We can also assume that $w_s$ is proportional to the inverse of the conditional expectation of the $s$-th term under the null hypothesis of independence (i.e. $w_s = [d_s(1 - d_s)]^{-1}$). $T_{CCS}$ subsumes then $T$ as a special case. Moreover, Nair (1987) showed that the distribution of $T$ can be approximated using the Satterthwaite's method (1946). See D'Ambra et al. (2018) for additional $T_{CCS}$ properties.

Beh et al. (2011) perform CA when cross-classified variables have an ordered structure by considering the Taguchi's statistic to determine graphically how similar (or not) cumulative response categories are with respect to criterion ones. This approach has been named "Cumulated Correspondence Analysis" (hereafter TCA). Let $\mathbf{W}$ be the $((J-1) \times (J-1))$ diagonal matrix of weights $w_j$ and $\mathbf{M}$ a $((J-1) \times J)$ lower unitriangular matrix of 1's. TCA amounts to the SVD$[\mathbf{D}_I^{\frac{1}{2}}(\mathbf{D}_I^{-1}\mathbf{P} - \mathbf{1}_I\mathbf{1}_J^T\mathbf{D}_J)\mathbf{M}^T\mathbf{W}^{\frac{1}{2}}] = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T$ with $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, such that

$$\frac{T_{CCS}}{n} = \text{trace}\,(\mathbf{D}_I^{\frac{1}{2}}(\mathbf{D}_I^{-1}\mathbf{P} - \mathbf{1}_I\mathbf{1}_J^T\mathbf{D}_J)\mathbf{M}^T\mathbf{W}\mathbf{M}(\mathbf{D}_I^{-1}\mathbf{P} - \mathbf{1}_I\mathbf{1}_J^T\mathbf{D}_J)^T\mathbf{D}_I^{\frac{1}{2}}) = \sum_{i=1}^{I} \lambda_i^2$$

Moreover, we highlight that above SVD is also equivalent to perform SVD of matrix $\mathbf{D}_I^{-\frac{1}{2}}(\mathbf{P} - \mathbf{D}_I\mathbf{1}\mathbf{1}^T\mathbf{D}_J)\mathbf{M}^T\mathbf{W}^{\frac{1}{2}}$. TCA decomposes then the Taguchi's statistic $T$ when $w_s$ is proportional to the inverse of the conditional expectation of the $s$-th term under the null hypothesis of independence ($w_s = [d_s(1 - d_s)]^{-1}$).

To visually summarize the association between the row and the column categories, TCA row and column principal coordinates are defined by $\mathbf{F} = \mathbf{D}_I^{-\frac{1}{2}}\mathbf{U}\boldsymbol{\Lambda}$ and $\mathbf{G} = \mathbf{W}^{-\frac{1}{2}}\mathbf{V}\boldsymbol{\Lambda}$, respectively. Here, $\mathbf{F}$ and $\mathbf{G}$ are matrices of order $I \times M$ and $(J-1) \times M$, respectively. The $s$-th row of matrix $\mathbf{G}$ contains the coordinates of category $y_{(1:s)}$ in the $M$ dimensional space (with $M = \text{rank}(\mathbf{D}_I^{\frac{1}{2}}(\mathbf{D}_I^{-1}\mathbf{P} - \mathbf{1}_I\mathbf{1}_J^T\mathbf{D}_J)\mathbf{M}^T\mathbf{W}^{\frac{1}{2}})$). Therefore, if there is approximately zero predicability of the column categories given the row categories then $\mathbf{F} \approx 0$ and $\mathbf{G} \approx 0$. To provide a more discriminating view of the difference between each cumulate rating category, authors consider also rescaling the row and column profile coordinates to obtain biplot-type coordinates (Goodman 1996): $\mathbf{F} = \mathbf{D}_I^{-\frac{1}{2}}\mathbf{U}\Lambda^{\alpha}$ and $\mathbf{G} = \mathbf{W}^{-\frac{1}{2}}\mathbf{V}\boldsymbol{\Lambda}^{(1-\alpha)}$ with $0 \leq \alpha \leq 1$. These coordinates are

related to the factorisation (for categorical data) proposed by Gabriel (Gabriel 1971) for the construction of the biplot.

Interested readers to this variant, which is linked with the partition of Taguchi's cumulative chi-squared statistic, can refer to (Beh et al. 2011; Sarnacchiaro and D'Ambra 2011; D'Ambra and Amenta 2011; D'Ambra et al. 2018) which discuss the technical and practical aspects of TCA in depth.

## 4 Restricted Cumulative Correspondence Analysis

Several authors (Böckenholt and Böckenholt 1990; Takane et al. 1991; Böckenholt and Takane 1994; Hwang and Takane 2002) pointed out that the interpretation of the multidimensional representation of the row and column categories may be simplified if additional information about the row and column structure of the table is available. Indeed, by incorporating this external information through linear constraints on the row and/or columns scores, a representation of the data may be obtained that is easier to understand and more parsimonious.

According to the principle of Restricted Eigenvalue Problem (Rao 1973), B&B (Böckenholt and Böckenholt 1990) proposed a canonical analysis of contingency tables which takes into account additional information about the row and column categories of the table. We name this approach "Restricted CA" (RCA). Additional information are provided in the forms of linear constraints on the row and column scores. Let $\mathbf{H}$ and $\mathbf{G}$ be the matrices of linear constraints of order $I \times E$ and $J \times L$ of ranks $E$ and $L$, respectively, such that $\mathbf{H}^T\mathbf{X} = \mathbf{0}$ and $\mathbf{G}^T\mathbf{Y} = \mathbf{0}$ where $\mathbf{X}$ and $\mathbf{Y}$ are the standardized row and column scores. RCA scores are obtained by a SVD of the matrix

$$\tilde{\boldsymbol{\Pi}} = \{\mathbf{I} - \mathbf{D}_I^{-\frac{1}{2}}\mathbf{H}(\mathbf{H}^T\mathbf{D}_I^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{D}_I^{-\frac{1}{2}}\}\boldsymbol{\Pi}\{\mathbf{I} - \mathbf{D}_J^{-\frac{1}{2}}\mathbf{G}(\mathbf{G}^T\mathbf{D}_J^{-1}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{D}_J^{-\frac{1}{2}}\}$$

that is $\tilde{\boldsymbol{\Pi}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T$ where $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix with eigenvalues $\lambda$ in descending order. Standardized row and column scores are given by $\mathbf{X} = \mathbf{D}_I^{-1/2}\mathbf{U}$ and $\mathbf{Y} = \mathbf{D}_J^{-1/2}\mathbf{V}$, respectively, such that $\mathbf{X}^T\mathbf{D}_I\mathbf{X} = \mathbf{I}$, and $\mathbf{Y}^T\mathbf{D}_J\mathbf{Y} = \mathbf{I}$, with $\mathbf{1}^T\mathbf{D}_I\mathbf{X} = \mathbf{0}$ and $\mathbf{1}^T\mathbf{D}_\mathbf{J}\mathbf{Y} = \mathbf{0}$. The classical approach to CA is obtained when $\mathbf{H} = (\mathbf{D}_I\mathbf{1})$ and $\mathbf{G} = (\mathbf{D}_J\mathbf{1})$ which represents the case of absence of linear constraints.

It is evident that, following the B&B's approach, we can also obtain an easier to understand and more parsimonious TCA graphical representation of the association between a nominal and an ordinal categorical variable. In this case we consider only additional information about the row (nominal) categories of the table. The additional information about the ordinal nature of the column variable is used by considering the cumulative sum of the cell frequencies across it. We use the same matrices of linear constraints $\mathbf{H}$ which ensures that the weighted average of the row TCA scores

equal zero. Restricted CA of cumulative frequencies (RTCA) amounts then to the SVD

$$[\mathbf{I} - \mathbf{D}_I^{-\frac{1}{2}}\mathbf{H}(\mathbf{H}^T\mathbf{D}_I^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{D}_I^{-\frac{1}{2}}]\boldsymbol{\Pi}_{(T)} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T \tag{3}$$

where $\boldsymbol{\Pi}_{(T)} = \mathbf{D}_I^{\frac{1}{2}}(\mathbf{D}_I^{-1}\mathbf{P} - \mathbf{1}_I\mathbf{1}_J^T\mathbf{D}_J)\mathbf{M}^T\mathbf{W}^{\frac{1}{2}}$ and such that $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. Standardized row and column RTCA scores are now given by $\mathbf{X} = \mathbf{D}_I^{-1/2}\mathbf{U}$ and $\mathbf{Y} = \mathbf{W}^{-\frac{1}{2}}\mathbf{V}$, respectively. TCA is then obtained when $\mathbf{H} = (\mathbf{D}_I\mathbf{1})$ which represents the case of absence of linear constraints. Note that single column categories are additionally plotted as supplementary points. Their column coordinates will be given by $\mathbf{Y}^+ = \boldsymbol{\Pi}_{(T)}^T\mathbf{U}$.

We point out that matrix $\mathbf{H}$ imposes the same constraints for all singular vectors $\mathbf{u}$ of SVD of identity (3), but it could be interesting to define different constraints on each singular vector. This aspect can be obtained by using a successive approach based on a rank-one reduction of the initial matrix $\boldsymbol{\Pi}_{(T)}$. Let $\boldsymbol{\Pi}_{(T)}^{(1)} = (\mathbf{I} - \mathbf{u}_1\mathbf{u}_1^T)\boldsymbol{\Pi}_{(T)}$ be the rank-one reduced matrix with respect to the first singular vector $\mathbf{u}_1$ corresponding to $\lambda_1$. This matrix is then substitute to $\boldsymbol{\Pi}_{(T)}$ in the SVD (3) and a new solution for $\mathbf{u}_2$, $\mathbf{v}_2$ and $\lambda_2$ is computed according to new linear constraints that we want to impose on this solution. New standardized row and column RTCA scores on this axis are so obtained. A new rank-one reduced matrix $\boldsymbol{\Pi}_{(T)}^{(2)}$ is then computed and substitute to $\boldsymbol{\Pi}_{(T)}$ in the SVD (3) for a solution with new constraints. New scores on $\mathbf{u}_2$ are consequently obtained and the approach is reiterated for the next axis $\mathbf{u}_3$, and so on. This iterative rank-one reduction approach is such that $\mathbf{X}^T\mathbf{D}_I\mathbf{X} = \mathbf{I}$ and $\mathbf{Y}^T\mathbf{W}\mathbf{Y} = \mathbf{I}$.

## 5 Example

In this section we illustrate the proposal method considering a data set (Table 1) mentioned in Agresti (2007).

The study is aimed at testing the effect of the factors urbanization and location on the ordered response preference for black olives of Armed Forces personnel. In particular, we have considered the case in which there is an asymmetric relationship between two categorical variables used as predictor variables (Urbanization and Region) and an ordinal response variable which is categorized into six ordered classes. The predictor variable Urbanization is characterized by two levels: Urban and Rural areas, whereas the predictor variable Region is characterized by three levels: North West, North East, and South West. The ordinal response variable is characterized by six growing ordered categories: A = dislike extremely, B = dislike moderately, C = dislike slightly, D = like slightly, E = like moderately, F = like extremely.

Since the ratings are ordered, a partition of Tagughi's inertia (of 0.1749) is applied by an unrestricted TCA which yields the singular values $\lambda_1 = 0.163$ and $\lambda_2 = 0.010$. TCA coordinates are displayed in Fig. 1. For the column categories, the label "(1)" reflects the "cumulative total" of rating 1 with those "(2:6)" of ratings 2, 3, 4, 5,

**Table 1** Data table

| | | Dislike extremity (1) | Dislike moderately (2) | Neither like nor dislike (3) | Like slightly (4) | Like moderately (5) | Like extremilly (6) | Total |
|---|---|---|---|---|---|---|---|---|
| Urban | North West | 20 | 15 | 12 | 17 | 16 | 28 | 108 |
| | North East | 18 | 17 | 18 | 18 | 6 | 25 | 102 |
| | South West | 12 | 9 | 23 | 21 | 19 | 30 | 114 |
| Rural | North West | 30 | 22 | 21 | 17 | 8 | 12 | 110 |
| | North East | 23 | 18 | 20 | 18 | 10 | 15 | 104 |
| | South West | 11 | 9 | 26 | 19 | 17 | 24 | 106 |
| Total | | 114 | 90 | 120 | 110 | 76 | 134 | 644 |

**Table 2** Eigenvalues

| TCA and RTCA Eigenvalues | | | | | | |
|---|---|---|---|---|---|---|
| Axis | Unconstrained | | | Constrained | | |
| | Eigenvalue | % | Cum. % | Eigenvalue | % | Cum. % |
| (1) | 0.163 | 93.206 | 93.206 | 0.070 | 95.480 | 95.480 |
| (2) | 0.010 | 5.587 | 98.793 | 0.003 | 3.822 | 99.302 |
| (3) | 0.002 | 1.169 | 99.962 | 0.001 | 0.698 | 100.000 |
| (4) | 0.000 | 0.027 | 99.989 | 0.000 | 0.000 | 100.000 |
| (5) | 0.000 | 0.010 | 100.000 | 0.000 | 0.000 | 100.000 |

and 6 given the Urban/Rural levels. Labels "(1:4)" and "(5:6)" reflect instead the comparison made of the cumulative total of ordered rating from 1 to 4 with 5 and 6, respectively, given the Urban/Rural levels. Similarly, labels "(1:3)" and "(4:6)" reflect the comparison made of the cumulative total of the ordered ratings from 1 to 3 with those of the remaining predictor categories, respectively, given the Urban/Rural levels. The remaining labels can be interpreted in a similar manner.

Note that Taguchi's analysis allows to identify how similar (or different) cumulate ordered column response categories are for each row category. Consider then Fig. 1 which graphically depicts about 98.79% of the association that exists between the two variables (see Table 2). It shows clearly that all the pairs of cumulated ratings are quite distinct, indicating that there is a perceived difference between these cumulate categories. The source of the variation between these ratings is dominated by all the Urban/Rural levels except for Urban North East (U.NE). The apparent difference between the most positive ratings and the others can be attributed mainly to U.NW and

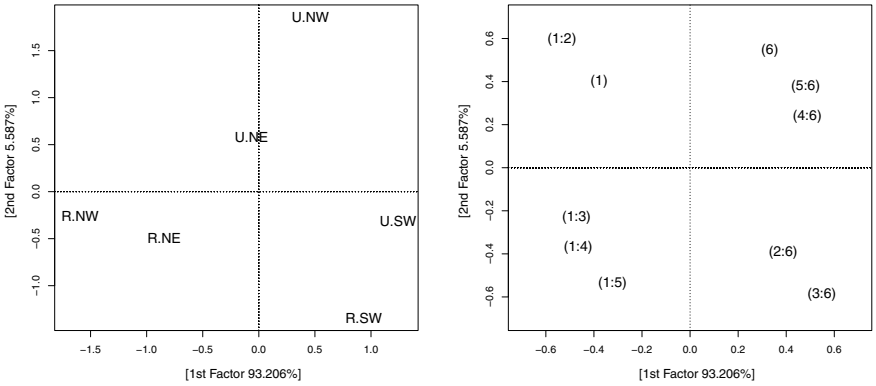



**Fig. 1** TCA plots

U.SW whereas the lowest values are characterized by R.NW and R.NE. First TCA axis depicts about 93.20% of global association and clearly contrasts the medium-low ratings with the high ones but there is not an evident separate domination by the Urban and Rural categories as well as by their three levels: North West, North East, and South West. Additional drawback of this plot is that it does not highlight the contribution of the high ratings "6" of U.NE (see Table 1). Indeed, it is a negligible category because the position of this point is closest to the origin.

In order to better highlight a contrast between the Urban and Rural categories on the first axis with respect to North West and South West levels, a RTCA solution is then computed by setting $\mathbf{H} = (\mathbf{D}_I \mathbf{1} | \mathbf{H}_R)$ with

$$\mathbf{H}_R^T = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}$$

RTCA plot (Fig. 2) now graphically depicts 99.30% of the global association and well highlights a contrast between the Urban and Rural categories as source of variation between the cumulated ratings. Rural categories are now source of the low ratings (left hands of Fig. 2), according to their frequencies of Table 1, as well as the Urban categories for the most positive ratings (right hands of Fig. 2). Moreover, this figure now points out a not negligible position of U.NE label showing the contribution of the high ratings "6" taken by this category.

We point out that introducing linear constraints in the TCA solution has then brought several advantages:

- a more parsimonious analysis is obtained where all the global association is depicted by only 3 axes (5 with TCA);
- an easier interpretation of the results is obtained highlighting a clear contrast between the Urban and Rural categories on the first axis as source of variation of the cumulated categories;

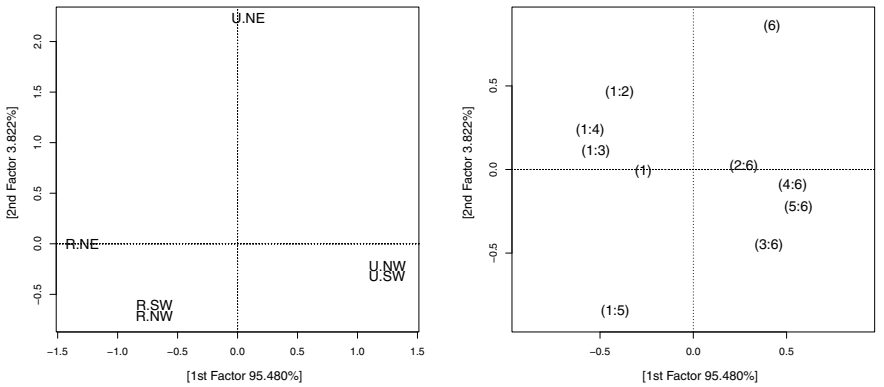



**Fig. 2** RTCA plots

- we observe an increase in the explained variability of the first axis: 95.48% of RTCA versus 93.21% of TCA;
- no predictive row category is now poorly represented.

## 6 Conclusion

Several authors highlighted that, introducing linear constraints on the row and column coordinates of a correspondence analysis representation, may greatly simplify the interpretation of the data matrix. Imposing also different constraints for each singular value may be useful in developing a parsimonious representation of a contingency table. B&B (Böckenholt and Böckenholt 1990) presented a generalized least squares approach for incorporating linear constraints on the standardized row and column scores obtained from a canonical analysis of a contingency table. This approach is based on the decomposition of a restricted version of the matrix of the Pearson ratios. Unfortunately the Pearson chi-squared statistic (likewise correspondence analysis) can perform poorly in studying the association between ordinal categorical variables (Agresti 2007). Beh et al. (2011) deal with this theme (Beh et al. 2011) by developing a CA extension (TCA) based on the Taguchi's statistic (Taguchi 1966, 1974). This statistic considers the cumulative frequency of cells in the contingency tables across the ordered variable and it has been introduced in the literature as simple alternative to Pearson's index for contingency tables with an ordered categorical variable.

A restricted extension of the Beh's approach (RTCA) has been here suggested to obtain a more parsimonious representation of the association and easier to explain. Natural forms of constraints may often appear from specific empirical questions asked by the researchers regarding the problem of their fields. In the exemplary application, introducing linear constraints in the TCA solution has brought several advantages in terms of interpretability and axes inertia rate. RTCA extends the

Cumulative Correspondence Analysis by taking into account external information (as linear constraints) and supplies a complementary interpretative enrichment of this technique as well as of the original CA approach.

## References


Agresti, A. (2007). *An introduction to categorical data analysis*. Wiley.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. (1972). *Statistical inference under order restrictions*. New York: Wiley.

Beh, E. (2004). Simple correspondence analysis: A bibliographic review. *International Statistical Review*, *72*(2), 257–284.

Beh, E. J., & Lombardo, R. (2012). A genealogy of correspondence analysis. *Australian & New Zealand Journal of Statistics*, *54*(2), 137–168.

Beh, E. J., & Lombardo, R. (2014). *Correspondence analysis: Theory*. Practice and New Strategies: Wiley.

Beh, E. J., D'Ambra, L., & Simonetti, B. (2011). Correspondence analysis of cumulative frequencies using a decomposition of Taguchi's statistic. *Communications in Statistics. Theory and Methods*, *40*, 1620–1632.

Benzecri, J. P. (1980). *Practique de l'analyse des donnees*. Paris: Dunod.

Böckenholt, U., & Böckenholt, I. (1990). Canonical analysis of contingency tables with linear constraints. *Psychometrika*, *55*, 633–639.

Böckenholt, U., & Takane, Y. (1994). Linear constraints in correspondence analysis. In: M. Greenacre & J. Blasius (Eds.), *Correspondence analysis in the social sciences: Recent developments and applications* (pp. 70–111). New York: Academic Press.

D'Ambra, A., & Amenta, P. (2011). Correspondence Analysis with linear constraints of ordinal cross-classifications. *Journal of Classification*, *28*, 1–23.

D'Ambra, L., Amenta, P., & D'Ambra, A. (2018). Decomposition of cumulative chi-squared statistics, with some new tools for their interpretation. *Statistical Methods and Applications*, *27*(2), 297–318.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*, 453–467.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. New York: Academic Press.

Goodman, L. A. (1996). A single general method for the analysis of cross-classified data: Reconciliation and synthesis of some methods of Pearson, Yule, and Fisher, and also some methods of correspondence analysis and association analysis. *Journal of the American Statistical Association*, *91*, 408–428.

Hwang, H., & Takane, Y. (2002). Generalized constrained multiple correspondence analysis. *Psychometrika*, *67*, 215–228.

Lebart, L., Morineau, A., & Warwick, K. M. (1984). *Multivariate descriptive statistical analysis: Correspondence analysis and related techniques for large matrices*. New York: Wiley.

Nair, V. N. (1986). Testing in industrial experiments with ordered categorical data. *Technometrics*, *28*(4), 283–291.

Nair, V. N. (1987). Chi-squared type tests for ordered alternatives in contingency tables. *Journal of American Statistical Association*, *82*, 283–291.

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press.

Rao, C. R. (1973). *Linear statistical inference and its applications*. Wiley.

Sarnacchiaro, P., & D'Ambra, A. (2011). Cumulative correspondence analysis to improve the public train transport. *Electronic Journal of Applied Statistical Analysis: Decision Support System and Services*, *2*, 15–24.

Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrical Bullettin*, *2*, 110–114.

Taguchi, G. (1966). *Statistical analysis*. Tokyo: Maruzen.

Taguchi, G. (1974). A new statistical analysis for clinical data, the accumulating analysis, in contrast with the chi-square test. *Saishin Igaku*, *29*, 806–813.

Takane, Y., Yanai, H., & Mayekawa, S. (1991). Relationships among several methods of linearly constrained correspondence analysis. *Psychometrika*, *56*, 667–684.

Takeuchi, K., & Hirotsu, C. (1982). The cumulative chi square method against ordered alternative in two-way contingency tables. Technical Report 29, Reports of Statistical Application Research. Japanese Union of Scientists and Engineers.

# Determining the Importance of Hotel Services by Using Transitivity Thresholds


**Pietro Amenta, Antonio Lucadamo, and Gabriella Marcarelli**

**Abstract** Customers' preferences related to the quality, the change, and the progress of their expectations have turned the quality in an indispensable competitive factor for hotel enterprises. The hotels have to evaluate the customer satisfaction and to assign to each factor a weight, expressing its importance for their customers. The aim of this paper is to evaluate the importance of hotel services. Our analysis involves more than 300 customers that answered to a survey and it takes into account five criteria: Food, Cleanliness, Staff, Price/benefit, and Comfort. To derive the ranking of preferences we used pairwise comparisons. The main issue linked to pairwise comparisons is the consistency of judgements. Transitivity thresholds recently proposed in literature give meaningful information about the reliability of the preferences. Our study shows how the use of ordinal threshold may provide a ranking of services different from that obtained by applying traditional consistency Saaty thresholds.

## 1 Introduction

Pairwise Comparison Matrices (PCMs) are widely used for representing preferences in multi-criteria decision problems. Given a set of $n$ elements, to derive the ranking of preferences by means of pairwise comparisons, a positive number $a_{ij}$ is assigned to each pair of elements $(x_i, x_j)$ with $i, j = 1, \ldots, n$. This number expresses how much $x_i$ is preferred to $x_j$ as regards a given criterion. By comparing all the elements, a positive square matrix $A = \left(a_{ij}\right)$ of order $n$ is then obtained. The value $a_{ij} > 1$ implies that $x_i$ is strictly preferred to $x_j$, whereas $a_{ij} < 1$ expresses the opposite preference, and $a_{ij} = 1$ means that $x_i$ and $x_j$ are indifferent (Saaty 1980, 1994). In order

P. Amenta · A. Lucadamo (✉) · G. Marcarelli
DEMM - University of Sannio, Piazza Arechi II, Benevento, Italy
e-mail: antonio.lucadamo@unisannio.it

P. Amenta
e-mail: amenta@unisannio.it

G. Marcarelli
e-mail: gabriella.marcarelli@unisannio.it



P. Mariani and M. Zenga (eds.), *Data Science and Social Research II*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-030-51222-4_3