DSSR 2024
Naples, 25th-27th March

Data Science & Social Research
4th International conference

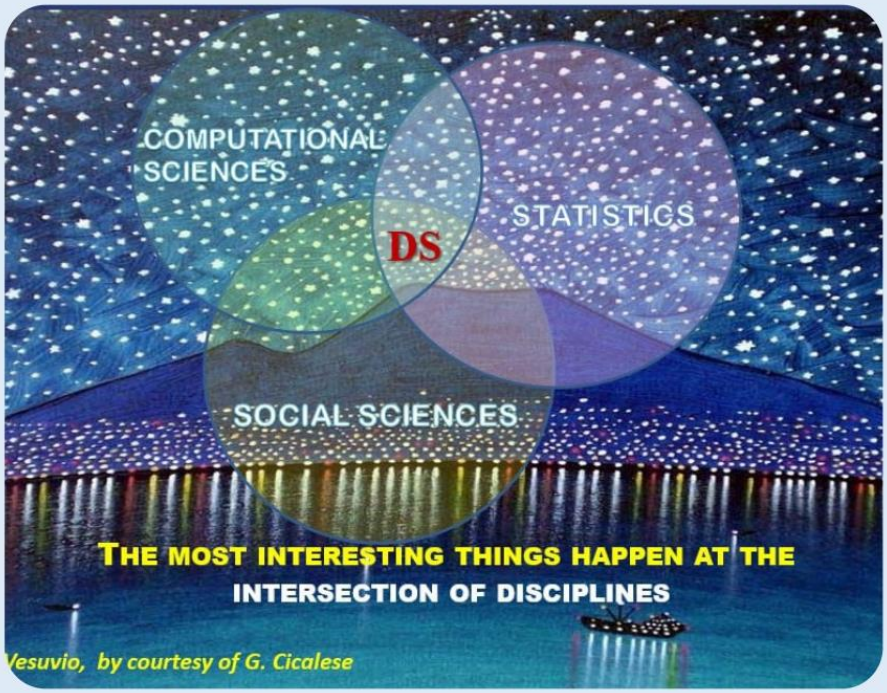
Abstract book

DSSR 2024

Data Science & Social Research

4th International conference

Naples, 25th-27th MARCH



Organised by
Department of Social Sciences (DISS) and
Department of Economics and Statistics (DISES)
University of Naples Federico II

PRIN 2022- The Future of Sustainability – P2022B3NFH – H53D23009470001-E53D23019030001

Plenary Session

Monday, March 25	
Enrico Giovannini	
Vincenzo Esposito Vinzi *Data & AI: for a future-fit Business Education*	
Tuesday, March 26	
Karina Gibert *The role of data and knowledge in the field of AI*	pag. 10
Sonia Stefanizzi *What kind of science is data science?*	pag. 11
Wednesday, March 27	
Ludovic Lebart *Visualization of Textual Data: some recent improvements*	pag. 12
Massimo Ragnedda *Bridging Theory to Practice: The Operationalization of Digital Capital*	pag. 13

Specialized and Contributed Session

1	Specialized session	**Epistemological and methodological issues in Digital social research**	Amaturo E., Masullo G.	1	Where is the love? The role of Algorithm Awareness in Tinder online dating	Felaco C., Acampa S.	pag. 15
				2	Exploring the Frontiers of Digital Social Research: Analysis of Distortions in Digital Data and Implications for Online Social Research	Lenzi F. R., Cavagnuolo M., Esposito V., Iazzetta F.	pag. 16
				3	Critical profiles of the use of algorithmic tools in the administration of justice	Pascali M.	pag. 17
				4	Digital-environmental habitus among Italian users: Using Path Structural Modelling to explore the role of Digital Expertise and Environmental Predispositions in enhancing digital sustainability	Ruiu M. L., Ragnedda M., Ruiu G.	pag. 18
2	Specialized session	**Populations and socio-demographic behaviors: some pieces**	Bonomo A., Strozza S.	1	Exploring Recent fertility trends in Italy by native and foreigners subgroups: A PLS Path Modeling Analysis	Mazza R., Pereiro T., Paterno A.	pag. 19
				2	The relationship between demographic dynamics and population ageing: a local multiscale approach	Benassi F., Buonomo A., Heins F., Strozza S.	pag. 20
				3	Political participation of immigrants and ethnic identity	Gatti R., Buonomo A., Strozza S.	pag. 21
				4	Sri Lankans in Italy: linking residential choices to spatial variations in the contextual socioeconomic conditions between and within eight main municipalities	Bitonti F., Benassi F., Mazza A., Strozza S.	pag. 22

Plenary Session

THE ROLE OF DATA AND KNOWLEDGE IN THE FIELD OF AI

Karina Gibert
Intelligent Data Science and Artificial Intelligence research center
e-mail: karina.gibert@upc.edu

The field of AI,born in 1956 is nowadays generating a lot of interest, since the maturity of ICT technologies enables the scalability of AI techniques to face the most complex problems we have ever met before. Climate change, sustainability, protein unfoldng are only some of the challenges that AI is able to tackle. In the talk we will travel from the origins of AI to the most recent advances by analysing not only the role of data structures and infrastructures, but also the role of specific domain knowledge and experts in the deployment of ethics, scalable and interpretable AI systemts. We will distinguish between symbolic and subsymbolic approaches, formal and informal management of knowledge, and the ethical framework proposed by the European Commission to frame the development of the field. Several real applications will help to understand the hipes and current challenges and we will provide main guidelines for a safe, trutsworthy and responsable AI.

WHAT KIND OF SCIENCE IS DATA SCIENCE?

Sonia Stefanizzi
University of Milano-Bicocca
e-mail: sonia.stefanizzi@unimib.it

The talk analyses the nature and role of data science, focusing on two main objectives: to provide a comprehensive definition and to criticise its relationship with other scientific disciplines. It begins by outlining the evolution of the concept of data in the context of the 21st century, highlighting the pioneering work of Jim Gray in proposing a new scientific paradigm based on data analysis. The comparison between data science and traditional scientific paradigms is discussed, identifying similarities and differences. The literature offers various definitions of data science, from minimalist to maximalist perspectives, highlighting the focus on prediction and information as fundamental components. The talk investigates whether data science is an established academic discipline or simply a set of pragmatic tools. Several views are considered, including Donoho's view that equates data science with statistics. The possibility that data science may be a transcendental discipline that supports other forms of research is also explored. The implications of 'agnostic science', which relies primarily on data to generate scientific knowledge, are analysed, and its integration with the traditional scientific method is discussed, emphasising the need for a balance between theory and data. Finally, we reflect on the need to develop frameworks to address epistemic and methodological challenges in the context of data science, especially regarding the transparency and interpretability of the algorithms used.

VISUALIZATION OF TEXTUAL DATA: SOME RECENT IMPROVEMENTS

Ludovic Lebart
Centre National de la Recherche Scientifique, Ecole Nationale Supérieure des Télécommunications of Paris
email: ludovic@lebart.fr

In the fields of Social Sciences, exploratory analyzes of textual data (EATD) have shown their usefulness in the case of the processing of responses to open-ended questions in large sample surveys or in the case of corpora of large texts (corpora of novels, political speeches, chronological textual series). Correspondence Analysis (CA) of lexical tables, with its simultaneous graphical representations of texts and words, is still often used together with clustering techniques. But the dimensionality and sphericity of the cloud of points may not allow for satisfactory visualizations in a low-dimensional space. The computation of additive trees (or phylogenetic trees) (1) is then essential. These trees produce 2- dimensional visualizations of high-dimensional spaces in which the real distances between elements in the full space can be easily inferred from clear rules of interpretation. In this contribution, we propose a new procedure for the simultaneous representation of texts and words in the framework of additive trees which allows us to combine the advantages of both CA and clustering. The trees drawn by certain representation algorithms (unrooted trees, such as, for example, the forcedirected graph drawings (2)) will be able to be enriched by a plot of the words used to construct them. We designate these simultaneous plots by the acronym ACT (for: "Additive Christmas Trees" … to remind that the additive trees whose vertices are the texts are "adorned" by the words of these texts – and more generally for a contingency or a binary table, e.g.: the tree of columns-points is enriched by the locations of the rows-points). The examples shown in this contribution, from political discourses to sets of poems, should exemplify the vital role of visualizations in a knowledge process. Meanwhile, we insist on the complementarity between AI tools and EATD.

References

(1) Huson D.H. and Bryant D. (2006). Application of Phylogenetic Networks in Evolutionary Studies, Molecular Biology and Evolution, vol. (23), 2: 254-267.
(2) Tutte, W. T. (1963), How to draw a graph, Proceedings of the London Mathematical Society, 13 (52): 743–768.

BRIDGING THEORY TO PRACTICE: THE OPERATIONALIZATION OF DIGITAL CAPITAL

Massimo Ragnedda
Northumbria University
email: massimo.ragnedda@northumbria.ac.uk

This presentation explores the translation of the concept of Digital Capital into practical application. Our central inquiries have revolved around defining Digital Capital and devising methods for its measurement. These questions have steered our research process, involving conceptualization, mapping, scale creation, concept validation, and real-world application. Drawing upon Pierre Bourdieu's theoretical framework, we have developed a conceptual understanding of how digital capital influences social inequalities and offers a framework to examine resource distribution, power dynamics, and interaction patterns within the digital sphere. Digital Capital represents a distinct form of capital that is amenable to empirical measurement. The focus of this presentation is on the mapping and creation of a scale for measuring Digital Capital, with a discussion on the introduction of this new theoretical concept and the methodological challenges inherent in its operationalization.

Specialized and Contributed Session

WHERE IS THE LOVE? THE ROLE OF ALGORITHM AWARENESS IN TINDER ONLINE DATING

Cristiano Felaco[1], Suania Acampa[2]

[1] *University of Naples, Federico II* (email: cristiano.felaco@unina.it)
[2] *University of Naples, Federico II* (email: cristiano.felaco@unina.it)

Algorithms are deeply ingrained in data curation processes, influencing decision-making across various sectors, including finance, healthcare, and education. The pervasiveness of algorithms in society has undoubtedly brought about positive consequences in enhancing efficiency, productivity, and convenience; however, algorithms can potentially harm specific individuals or social groups. Therefore, being aware that algorithms are used in online applications, for what purposes, and in which online contexts (1;2), as well as being able to interact with them, is considered a fundamental digital requirement (3). In contrast, a lack of or limited algorithm awareness makes individuals unknowingly reliant on these algorithmic configurations. Considering these aspects, the study investigates the extent of user awareness concerning the algorithmic mechanisms governing the online dating platform Tinder. Specifically, we interviewed 20 Active Tinder users to understand Tinder's algorithmic processes and how such awareness (or presumed) might influence how they approach dating. The primary outcomes suggest that a subset of the respondents acquire knowledge about the operational mechanisms of the dating platform either through firsthand experience (4) or via information exchange (5). Diverse experiences precipitate a spectrum of responses that influence user interaction with these platforms, occasionally driving attempts to circumvent the underlying algorithmic principles. Individuals with a profound comprehension of Tinder's algorithmic underpinnings implement various tactics and strategies, characterized as "bottom-up," to sway the results to their advantage. Such strategies encompass profile optimization and a targeted approach to swiping. By employing these techniques, users can markedly bolster their visibility and the likelihood of establishing substantive connections on the platform. A different awareness of the algorithmic system may give users a distinct advantage in exploiting information about how these platforms rate and display profiles to other users.

References

(1) Hamilton, K., Karahalios, K., Sandvig, C., & Eslami, M. (2014). A path to understanding the effects of algorithm awareness. CHI'14 Extended Abstracts on Human Factors in Computing Systems, 631– 642.
(2) Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). I always assumed that I wasn't really that close to [her]. In B. Begole, J. Kim, K. Inkpen, & W. Woo (Chairs), The 33rd annual ACM conference, Seoul, Republic of Koreaz
(3) Gran A-B., Booth P., Bucher T. (2021). To be or not to be algorithm aware: a question of a new digital divide? Information, Communication & Society, n. 24, pp. 1779-1796, 2021.
(4) Cotter, K. (2019). Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram. New Media & Society, 21(4), 895-913.
(5) Bishop S. (2019). Managing visibility on YouTube through algorithmic gossip. New media & Society, n.31, pp. 2589-2606, 2019.

EXPLORING THE FRONTIERS OF DIGITAL SOCIAL RESEARCH: ANALYSIS OF DISTORTIONS IN DIGITAL DATA AND IMPLICATIONS FOR ONLINE SOCIAL RESEARCH

Francesca Romana Lenzi[1], **Michela Cavagnuolo**[2], **Vincenzo Esposito**[3], **Ferdinando Iazzetta**[4]

[1] *University of Roma "Foro italico"* (email: francescaromana.lenzi@uniroma4)
[2] *University of Roma "Foro italico"* (email: michela.cavagnuolo@uniroma1.it)
[3] *University of Roma "Sapienza"* (email: vi.esposito@uniroma1.it)
[4] *University of Roma "Sapienza"* (email: ferdinando.iazzetta@uniroma1.it)

The present study titled "Exploring the Frontiers of Digital Social Research: Analysis of Distortions in Digital Data and Implications for Online Social Research" delves into the analysis of distortions that occur in digital data (1) and their implications for online social research (2). Due to the pervasiveness of digital technologies and the growing role of online communication channels in our everyday lives, researchers have increasingly turned to digital data sources to study social phenomena. However, the accuracy and validity of such data are often compromised by various distortions, posing significant challenges to conducting reliable research. The research aims to analyze studies in sociology that use digital data and have questioned the quality of the data. Using the framework by Olteanu et.al (3) that describes biases and pitfalls while working with social data, the study identifies various sources of bias in digital data such as sampling bias, measurement error, and selection bias. The analysis involves an extensive review of existing literature and empirical data. From the empirical point of view, through a literature review, we want to extrapolate and study articles in sociology using digital data published on the Web of Science indexing platform. The selection criteria for article extraction are 1. Field of study: (Sociology) 2. Publication type (Articles); 3. Language (English); 4. Publication format (Open Access). The words used for extraction are Digital data, Data Quality, Digital traces, Distortion, Reliability, Validity, Generalizability, Trustworthiness, Credibility, Transferability, Confirmability, Biases, Evaluation, and Ethics for a total of 129 articles then subjected to quality control by the researchers to identify the articles most relevant to the study.

References

(1) Rogers, R. (2015). Digital methods for web research. Emerging trends in the social and behavioral sciences:An interdisciplinary, searchable, and linkable resource, 1-22.
(2) Delli Paoli, A., Masullo, G. (2022). Digital Social Research: Topics and Methods. [Italian Sociological Review,12 (7S), 617-633]
(3) Olteanu, A., Castillo, C., Diaz, F., and Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Front. Big Data 2:13.

CRITICAL PROFILES OF THE USE OF ALGORITHMIC TOOLS IN THE ADMINISTRATION OF JUSTICE

Michelangelo Pascali

University of Naples, Federico II (email: Michelangelo.pascali@unina.it)

The use of decision-making automation tools is increasingly prevalent in the administration of justice. This is due to the problematic nature of human decision-making in regard to ascertainment and forecasting activities. Technological evolution could therefore provide new, partially algorithmic solutions to address the social risks associated with inadequate management of the issues at hand, especially in relation to criminal phenomena. However, it is important to acknowledge that this may have its own limitations and challenges to overcome.

DIGITAL-ENVIRONMENTAL HABITUS AMONG ITALIAN USERS: USING PATH STRUCTURAL MODELLING TO EXPLORE THE ROLE OF DIGITAL EXPERTISE AND ENVIRONMENTAL PREDISPOSITIONS IN ENHANCING DIGITAL SUSTAINABILITY

Maria Laura Ruiu[1], **Massimo Ragnedda**[2], **Felice Addeo**[3], **Gabriele Ruiu**[4]

[1] Northumbria University (email: maria.ruiu@northumbria.ac.uk)
[2] Northumbria University (email: massimo.ragnedda@northumbria.ac.uk)
[3] Univeristy of Salerno (email: faddeo@unisa.it)
[4] University of Sassari (email: gruiu@uniss.it)

This study explores how environmental predispositions, backgrounds, and digital expertise influence digital behaviours and pro-environmental awareness among ICT users in Italy. Thedigital-environmental habitus, reflecting digital technology use and environmental attitudes, is explored through a survey of 1188 participants. Ruiu et al. (2023) define the digital-environmental dimension of the habitus as encompassing pre-existing backgrounds, as well as the assimilated increased use of digital technologies in people's daily lives, reflecting individual environmental attitudes. The study employs this concept to understand how individuals respond tothe challenges presented by digital advancements in times of environmental crisis. It also advances the concept by breaking it down its two constituent dimensions: awareness and engagement. Analysing these dimensions independently enabled us to study how individuals' offline environmental values are mirrored in digital experiences. Results from a Path Structural Model indicate that environmental predispositions significantly impact digital pro-environmental awareness and behaviours. The existence of digital-specific environmental awareness enhances pro-environmental digital behaviours, emphasising the importance of raising awareness about the environmental impact of digital tools. While digital expertise alone does not significantly predict digital environmental awareness, it does moderate the behavioural aspect of the digital environmental habitus, promoting behaviours that are mutually beneficial for users and the environment. Therefore, digital skills might have the potential to amplify behaviours driven by cognitive and technical-based mechanisms, as suggested by the literature. These mechanisms can empower individuals to effectively use digital tools for retrieving and critically analysing environmental information in a manner that benefits both themselves and the environment. This is in line with studies that emphasise how ICTs may facilitate access to up-to-date scientific knowledge, and support different cognitive processes (Al-Ansi, Garad, and Al- Ansi, 2021), including knowledge acquisition and problem-solving (Muzana et al., 2021). These findings highlight the importance of promoting environmental awareness among digital users. It also emphasises the role of digital expertise in shaping digital behaviours.

EXPLORING RECENT FERTILITY TRENDS IN ITALY BY NATIVE AND FOREIGNERS SUBGROUPS: A PLS PATH MODELING ANALYSIS

Rocco Mazza[1], Thais Garcia Pereiro[2], Anna Paterno[3]

[1] *Università degli studi di Bari Aldo Moro* (email: rocco.mazza@uniba.it)
[2] *Thais Garcia Pereiro*, Università degli studi di Bari Aldo Moro (email: t-garcia.pereiro@uniba.it)
[3] *Anna Paterno*, Università degli studi di Bari Aldo Moro (email: anna.paterno@uniba.it)

Over the past decades, European countries have been facing a systematic decline in rates of natural population change, counteracted by positive net international migration rates driving overall population growth. International migration might play a dual role, first, by directly contributing to population change and indirectly, influencing demographic age and sex structures, changing its composition. Within the broader European scenario, Italy's distinctive demographic challenges, marked by enduringly low fertility trends, sustained international migration, and accelerated population ageing, warrant special attention. This paper aims to investigate the impact of determinants on fertility by Italian population subgroups of foreigners and natives. Utilizing demographic macro-data from ISTAT, our analysis employs the Partial Least Squares (PLS) path modelling technique to model the socio-economic factors that have the greatest impact on the fertility of the groups studied. This methodological framework enhances our comprehension of the intricate dynamics governing the demographic landscape, offering insights into the unique role of international migration in shaping fertility trends in Italy.

THE RELATIONSHIP BETWEEN DEMOGRAPHIC DYNAMICS AND POPULATION AGEING: A LOCAL MULTISCALE APPROACH

Benassi Federico[1], Buonomo Annamaria[2], Frank Heins[3], Salvatore Strozza[4]

[1] *University of Naples Federico II* (email: federico.benassi@unina.it)
[2] *University of Naples Federico II* (email: annamaria.buonomo@unina.it)
[3] *National Research Council, Institute of Research on Population and Social Policies* (email: frank.heins@bergen-online.de)
[4] *University of Naples Federico II* (email: salvatore.strozza@unina.it)

The aim of the contribution is to identify the relationship between demographic dynamics and population ageing. Both processes have important geographical specificities that are amplified on a local scale. The population of Italy is affected by a widespread ageing process and will experience a significant demographic contraction. However, not all municipalities are ageing at the same rate and with the same intensity and not all will face demographic shrinking. The contribution intends to study the links between these two processes considering spatial heterogeneity from a local multiscale perspective. To this end, indicators of population ageing will be placed in relation to a series of indicators relating to demographic dynamics. All indicators are calculated at municipal level and refer to the last twenty years. The analysis methods include classical (i.e. global non-spatial) and local multiscale geographically weighted regression models. The results will allow to appreciate not only the relationships between demographic dynamics and population ageing but, above all, how these relations change geographically, at what scale they operate and what type of geographical patterns they draw. These three aspects might be of particular interest to calibrate place-based policies.

IMMIGRANT POLITICAL PARTICIPATION AND ETHNIC IDENTITY

Rosa Gatti[1], Anna Maria Buonomo[2], Salvatore Strozza[3]

[1] *University of Naples Federico II* (email: rosa.gatti@unina.it)
[2] *University of Naples Federico II* (email: annamaria.buonomo@unina.it)
[3] *University of Naples Federico II* (email: salvatore.strozza@unina.it)

International scholarship is increasingly addressing the determinants of political engagement among individuals with migratory backgrounds. Only a few studies have investigated the role played jointly by ethnic identity and perceived discrimination in influencing political engagement. No studies have considered differences by migratory generation. This contribution tries to fill this gap using data from the Social Condition and Integration of Foreign Citizens survey of the Italian National Institute of Statistics (2011–2012). In the proposed analyses, we measured political engagement using three different variables, two for attitudinal political engagement (interested in Italian politics and talking about politics) and one for behavioural political engagement (participating in political debate). We applied a set of logistic regressions with findings presented as average marginal effects, and we deepened the results by applying the interaction between ethnic identity and perceived discrimination. The empirical results indicate that for both the first generation and 1.5 generation, there is no evidence of a negative role played by the preservation of the minority identity as long as it is also accompanied by the acquisition of a majority (national) identity. Furthermore, the results proved that as discrimination increases, political engagement (both attitudinal and behavioural) increases for both migrant generations considered.

SRI LANKANS IN ITALY: LINKING RESIDENTIAL CHOICES TO SPATIAL VARIATIONS IN THE CONTEXTUAL SOCIOECONOMIC CONDITIONS BETWEEN AND WITHIN EIGHT MAIN MUNICIPALITIES

Francesca Bitonti[1], Federico Benassi[2], Angelo Mazza[3], Salvatore Strozza[4]

[1] *University of Catania* (email: francesca.bitonti@unict.it)
[2] *University of Naples Federico II* (email: federico.benassi@unina.it)
[3] *University of Catania* (email: angelo.mazza@unict.it)
[4] *University of Naples Federico II* (email:salvatore.strozza@unina.it)

Summary of background data: The current study proposes a comparative spatial analysis of residential segregation and settlement models among Sri Lankans in the eight main Italian municipalities, where the majority of the community resides. Objectives: The research serves a dual purpose based on the geographic scale considered, encompassing both inter-urban and intra-urban levels of analysis. Firstly, it seeks to compare the patterns of Sri Lankan settlements across diverse urban contexts. Secondly, it endeavors to identify potential spatial polarization of Sri Lankans within specific neighborhoods, examining spatial correlations with key variables indicative of socioeconomic disparities in urban areas. Methods: The initial phase involves generating descriptive statistics and maps to preliminarily examine the distribution of Sri Lankans in each municipality under consideration. Subsequently, multiple linear models are employed to evaluate the overall variation in the concentration of Sri Lankans concerning various socioeconomic predictors. Additionally, geographically weighted regressions are implemented to explicitly model spatial dependence between Sri Lankans' location quotients and the specified predictors. All variables are referenced to a common geographic grid, enabling the standardization of different areal unit arrangements and facilitating spatial comparisons. Results: The findings reveal that the distinctive residential patterns of Sri Lankans extend beyond a simplistic center-periphery dichotomy. Discussions/conclusions: The ethnic mixing observed in historic centers may indeed unveil states of socioeconomic inequality, necessitating tailored interventions and thoughtful consideration.

DO ALGORITHMS DREAM OF DIGITAL SOCIETIES? EXPLORING HUMAN AND AI INTERACTIONS IN THE DATA AGE

Edmondo Grassi

Università Telematica San Raffaele Roma (email: edmondo.grassi@uniroma5.it)

Investigating the formation of algorithmic identities, their peculiarities, and the relational modes developed by humans with them are fundamental themes for understanding the transformative changes affecting contemporary societies and cultures. The presence of intelligent algorithms, intangible and robotic, overt and latent, supportive and retroactive, influences how individuals articulate their values, behaviour patterns, ethical principles, beliefs, and connections, all the way to the identity project of who they are and who they aspire to become. Among the challenges and discoveries that this framework entails in the field of research, there arises the need to understand who the future "subjects of study" will be. From an STS perspective, it aims to provide a trans-speciesism reflection since these agents not only influence human behaviour but are also integral to the generation and collection of digital data. The intent is to elucidate their Manifestations and their hypothetical degrees of influence and interaction. Furthermore, the analysis of algorithmic intelligent agents could offer critical insights into digital research methods, particularly in terms of ethics, transparency, and awareness, both on the part of researchers and the subjects involved, raising relevant questions about how these entities influence, manipulate, distort, and accompany human interactions, and how this translates into the reconfiguration of collected data. The objectives of the presentation will, therefore, be to delineate algorithmic agents and their social manifestations and to reflect on the symbolic aspects adhering to anthropocentric social paradigms, determining further evolutionary processes.

CROWDSOURCING PLATFORMS IN DIGITAL SOCIAL RESEARCH: METHODOLOGICAL AND ETHICAL ISSUES

Silvia Cataldi[1], Maria Carmela Catone[2]

[1] *University of Rome* "La Sapienza" (email: silvia.cataldi@uniroma1.it)
[2] *University of Salerno* (email: mcatone@unisa.it)

The latest frontiers in digital social research often involve the use of crowdsourcing platforms to support the implementation of different empirical studies. These are online tools which combines traditional modes with those linked to the developments in digitization processes; they are especially used to quickly and cheaply recruit participants and collect data, carry out experiments, validate research instruments such as questionnaires, and explore a wide range of social phenomena, enabling researchers to obtain data from participants from different cultures and backgrounds (1,2). In this contribution, starting with an overview of major crowdsourcing platforms like Amazon's Mechanical Turk and Prolific, the key features of this strategy for conducting social research will be described. Specifically, methodological and ethical issues related to data quality and validity, sample representativeness, and informed consent of the participants will be addressed. For example, the heterogeneous composition of the online population involved may lead to unrepresentative samples, with the possibility of bias in the results; the flexibility of crowdsourcing platforms can make it to control the conditions under which the study is carried out, affecting the validity of the data collected. From an ethical point of view, there are issues related to the compensation of participants that might encourage the search for quick solutions at the expense of the quality of the data collected.

References

(1) Gleibs, I. H. (2017). Are all "research fields" equal? Rethinking practice for the use of data from crowdsourcing market places. Behavior Research Methods, 49(4), 1333-1342.
(2) Zhao, Y., & Zhu, Q. (2014). Evaluation on crowdsourcing research: Current status and future direction. Information systems frontiers, 16, 417-434.

GEO-MEDIA AND SOCIAL STRATIFICATION: A CASE STUDY

Ciro Clemente De Falco[1], Emilia Romeo[2], Antonio De Falco[3], Marco Ferracci[4]

[1] *University of Naples Federico II* (email: ciroclemente.defalco@unina.it)
[2] *University of Salerno* (email: eromeo@unisa.it)
[3] *University of Milan Bicocca* (email: antonio.defalco@unimib)
[4] *University of Naples Federico II* (email: marco.ferracci@unina.it)

Data volume, variety, and velocity from various sources have increased dramatically in recent years. Most of this big data are produced by web users: the so-called user-generated content (UGC). This type of data can include geographic information provided voluntarily by users through the geographic localisations of their devices. Twitter (now called X) is a primary source of geodata for social research. Through a case study, this research aims to understand the spatial representativeness of geodata by seeking whether the spatial distribution of geoTwitter is related to the infrastructural characteristics of the area (e.g., museums, parks, B&Bs, historical sites, etc.) or its socioeconomic characteristics. To address this research objective, we collected 100,000 geolocalised tweets in the city of Naples using a Python scraper. Data were collected between January 1, 2020, and March 31, 202. Geolocalised tweets were analysed with socio-economic data from the 2021 census and leisure facilities and services of the area. The analysis results show that geolocated tweets are more closely related to the physical structures within the area rather than its socioeconomic conditions.

LIMITATIONS AND OPPORTUNITIES OF SOCIAL RESEARCH ON X. A STUDY ON THE ITALIAN CASE OF CHATGPT-4

Caterina Ambrosio[1], Vincenzo Laezza[2], Ciro Clemente De Falco[3]

[1] *University of Naples Federico II*
[2] *University of Naples Federico II* (email: vincenzo.laezza@unina.it)
[3] *University of Naples Federico II* (email: ciroclemente.defalco@unina.it)

X, formerly known as Twitter, has played a fundamental role in social research, emerging as the primary platform for numerous scholarly investigations. However, recent scholarly discourse has increasingly underscored critical points, such as the absence of geographical information and the closure of APIs. It is within this context that this present study is positioned; its primary objective is to scrutinize the current efficacy of X as a source of big data for social scientists. The chosen case study focuses on Italian-language posts from March 31st to April 28th, during which Italians were barred from accessing ChatGPT due to the intervention of the privacy regulator. The empirical phase brought to light challenges in the platform's utilization. Foremost among these challenges was data collection. Following the closure of APIs, which presented a set of complications, data collection was executed through an algorithm operating on both the web page and HTML structure. While this method allows for a greater volume of records compared to previous approaches, it comes at the cost of significantly diminished informational quality. The prevalence of null values and the almost complete absence of metadata are illustrative examples. The principal consequence is the restriction to a limited set of analysis techniques, primarily constricted to text-based analysis. Furthermore, another salient aspect to consider is the increasingly prominent role of images in X communication, prompting the imperative to develop tools capable of capturing and analyzing such elements, with contributions from visual sociology. In conclusion, this work, grounded in a case study, elucidates critical issues and strengths in utilizing X as a resource for social research.

THE EVOLUTION OF SOCIAL FRAILTY IN SOCIAL DOMAIN: A BIBLIOMETRIC ANALYSIS

Giulia Cavrini[1], Maria Gabriella Grassia[2], Marina Marino[3], Agostino Stavolo[4]

[1] *Free University of Bozen* (email: gcavrini@unibz.it)
[2] *University of Naples Federico II* (email:mariagabriella.grassia@unina.it)
[3] *University of Naples Federico II* (email:marina.marino@unina.it)
[4] *University of Naples Federico II* (email: agostino.stavolo@unina.it)

The increasing aging population has sparked heightened interest in topics like frailty and life satisfaction. One often-neglected aspect is social frailty, characterized by insufficient engagement in social networks and a perceived lack of contact and support (1). Despite its significance, social frailty remains among the least explored areas in frailty literature. To gain a comprehensive understanding of this subject, it is crucial to review and map existing literature. Scientific knowledge mapping involves a thorough examination of relevant articles. This study specifically focuses on science mapping, enabling the identification and visualization of themes and trends both synchronically (within a specific period) and diachronically (across time) (2). The objective is to trace the evolution of themes related to social frailty in social fields using thematic map evolution. Using the query TS=(("social frailt*" OR "social vulnerabilit*") AND ("elderl*" OR "age*" OR "old*" OR "old* adult*")), we analysed 370 documents on social frailty from 1997 to 2023 in the Web of Science (WoS). These documents span various categories (Anthropology, Communication, Demography, Development Studies, Family Studies, Gerontology, Multidisciplinary Sciences, Women's Studies, Psychology (Clinical, Applied, Developmental, Educational, Multidisciplinary, Social), Social Issues, Social Work, Interdisciplinary Social Sciences, Educational Research, and Health Policy Services). The analysis tracks the evolution of research themes in scientific production on social frailty. Initially, the focus cantered on physical and disease aspects, particularly Alzheimer's and dementia. This shifted towards investigating the impact of climate change on older adults, followed by a concentration on the social effects of the pandemic. In the last two years, there has been a notable emphasis on gender inequalities and social support. This progression underscores a broader recognition of environmental, societal, and gender-related dimensions in understanding frailty and aging, reflecting a more comprehensive research approach in this field.

References

(1) Bunt, S., Steverink, N., Olthof, J., Van Der Schans, C. P., & Hobbelen, J. S. M. (2017). Social frailty in older 1adults: a scoping review. European journal of ageing, 14, 323-334.
(2) Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. Journal of Informetrics, 5(1), 146–166.

THE STUDENT POPULATION WITH MIGRATORY BACKGROUND: STATISTICAL CHALLENGES OF A SELF-SELECTED SAMPLE

Lorenzo Giammei[1], Laura Terzera[2], Fulvia Mecatti[3]

[1] *Italian National Research* Council (email: lorenzo.giammei@unimib.it)

[2] *University of Milan-Bicocca* (email: laura.terzera@unimib.it)

[3] *University of Milan-Bicocca* (email: fulvia.mecatti@unimib.it)

The University of Milan-Bicocca conducted a survey to study its student population with a migratory background. However, the interpretation of the results must consider the non-probability nature of the employed sample, which may be affected by self-selection, potentially leading to biased descriptive statistics. The University also gathers yearly administrative data which unfortunately do not provide enough information to map the target population. Survey and administrative data can be matched through a unique linkage key. The main objective of this work lies in mapping the entire target population within the administrative data. In particular we aim at investigating if each student belongs or not to the target population and its migratory background type. We exploit survey data to map the hidden sub-population. The survey in fact contains an initial set of screening questions that has a two-fold implication: it directly allows to identify, among the respondents of the survey, individuals belonging to the target population; it indirectly allows to keep track of all the respondents that do not belong to the target population, since the meta-information of discarded interviews remains stored in the platform. We employ this information, that is missing in the administrative dataset, to fit a model that predicts if a student belongs or not to the target population, for all the units in the administrative data. The model allows mapping the target population through a prediction process. The robustness and accuracy of the prediction is evaluated by employing both statistical and machine learning models and through k-fold cross-validation. The identification of the hidden sub-group enables the study of its characteristics, which would have been impossible without resorting to a predictive model. Possible improvements and extensions of the proposed methodology to other applications that share a similar data structure are presented.

MISCONCEPTIONS OF SOCIAL POSITIONING ACROSS TIME AND SPACE

Daniela Bellani[1], Eleonora Clerici[2], Nevena Kulic[3], Debora Mantovani[4], Loris Bergolini[5]

[1]*Milan Catholic University of the Sacred Heart* (email: daniela.bellani@unicatt.it)
[2] *University of Pavia* (email: eleonora.clerici@unipv.it)
[3] *University of Pavia* (email: nevena.kulic@ unipv.it)
[4] *University of Bologna* (email: d.mantovani@unibo.it)
[5] *University of Bologna* (email: loris.vergolini@unibo.it)

Our contribution aims at understanding inequality (mis)perceptions and discrepancies in Europe, focusing on social classes. Generally speaking, scholars have focused on objective social positioning (in terms of income or occupational positioning) in order to explain several individual behaviors, such as supports for political parties, redistributive policies, welfare state (2). However, a very recent literature questions this approach, claiming that misconceptions of social positioning matter for demand for policies oriented toward redistribution (1). It is interesting to report that, in certain countries, middle class individuals misperceive their social positioning more than other social groups (3). We exploit five ad hoc modules of ISSP (International Social Survey Programme) on Social Inequality. These modules fit with our goals given that they deal with issues such as views on earnings and incomes, career advancement by means of family background and networks, social cleavages and the social position of the individuals and their partner. Implementing multilevel regression models and adopting a novelty approach based on both individual and partner information, this contribution sheds new light on the determinants of misperception of social positioning.

References

(1) Bussolo, M., Ferrer-i-Carbonell, A., Giolbas, A., & Torre, I. (2021). I perceive therefore I demand: The formation of inequality perceptions and demand for redistribution. Review of Income and Wealth, 67(4), 835-871.
(2) Knell, M., & Stix, H. (2020). Perceptions of inequality. European Journal of Political Economy, 65, 101927.
(3) Gimpelson, V., & Treisman, D. (2018). Misperceiving inequality. Economics & Politics, 30(1), 27-54.

UNVEILING THE POWER OF PUBLICWORKSFINANCEIT R PACKAGE FOR ANALYZING AND VISUALIZING ITALIAN SOIL DEFENSE INVESTMENTS

Lorena Ricciotti[1], Alessio Pollice[2]

[1] *University of Bari Aldo Moro* (email: lorena.ricciotti@uniba.it)
[2] *University of Bari Aldo Moro* (email: alessio.pollice@uniba.it)

The package "PublicWorksFinanceIT", enables users to retrieve and analyze financial data related to public works in Italy, specifically, it focuses on soil defense investments. The data is sourced from three distinct platforms: the OpenCoesione repository, which aggregates financial information for public works founded national and European funds, the Ministry of Economy and Finance's OpenBDAP open data platform, housing all Italian funds allocated to public interventions, and the ReNDiS database, provided by ISPRA, that exclusively gathers information about interventions in soil defense. This package offers a user-friendly tool that eliminates the need for direct access to the aforementioned institutional platforms and ensures real-time updates. Additionally, all measurements, metadata, and accompanying analytical tools are provided in English, enhancing accessibility for both international and domestic users. The data from these three sources are linked using two distinct variables: the unique project code (CUP) and the local project code, ensuring that there is no duplication of data. Moreover, the data is geographically referenced, meaning that each financial observation is associated with public work in a specific municipality within a particular Italian region. This includes information on the region, province, and municipality for each dataset entry. Indeed, the dataset has been enriched with geo-references for each municipality. Users can select the reference from either the coordinates of the municipality's centroid or the areal data of the polygon shape representing the municipality. In addition to functions for data retrieval, there are also functions available for visualizing the collected data on maps, classified by various variables.

THE USE OF MACHINE LEARNING TECHNIQUES ON THE INTEGRATE ADMINISTRATIVE DATA SYSTEM AIMS TO ENHANCE THE ACCURACY OF THE POPULATION CENSUS COUNT

Antonio Laureti Palma[1], Gerardo Gallo[2]

[1] *Italian National Institute of Statistics* (email: lauretip@istat.it)
[2] *Italian National Institute of Statistics* (email: gegallo@istat.it)

Since 2020, the population census count is carried out through the integrated use of administrative sources. The new census process enables the observation of individuals' Signs of Life (SoL) in terms of usual residence in Italy by integrating information on the place of residence recorded in the population register with that derived from administrative sources (e.g. Labor and Education archive, Tax Returns archive, Earnings, Retired, and Non-Pension Benefits archives, Permits to Stay archive, electricity and gas consumption archives, etc.). As is widely known, the population register is affected by under and over coverage, particularly for some population groups (e.g. foreigners). In addition, the "SoL" approach depends greatly on the location of the signals recorded in the administrative sources. In particular, place of residence discrepancies can cause misplacement errors that produce over and under-counting, simultaneously within two different municipalities. In this study, SoL are used to implement a ML classification strategy to distinguish between usually resident and not usually resident population in Italy. Supervised training and testing data sets are built using the 2021 census sample survey data. Different ML models and hyper-parameters are used to identify the best model for classification. In order to deepen the analysis of some quantitative differences between the place of registration of individuals in the population registers and the usual residence as traced by the administrative sources, an unsupervised analysis of electricity and gas consumption is developed. Through the analysis between the information on households included in statistical registers and the consumption patterns of the smart meters associated, we try to assess the most probable place of usual residence, i.e. where a household, or part of it, actually lives, reducing the possible misplacement errors.

References

(1) M. Zuppardo, V. Calian, Ó. Harðarson, "Machine learning methods for estimating the Census population", Nordic Statistical Meeting 2022.
(2) UNECE, Guidelines for Assessing the Quality of Administrative Sources for Use in Censuses, United Nations, New York,https://unece.org/sites/default/files/2021-10/ECECESSTAT20214_WEB.pdf (2021).

ISTAT NEW ENUMERATION AREAS 2021 FOR SPATIAL ANALYSIS

Stefano Mugnoli[1], Fabio Lipizzi[2], Alberto Sabbi[3]

[1] *Italian National Institute of Statistics* (email: mugnoli@istat.it)
[2] *Italian National Institute of Statistics* (email: fabio.lipizzi@istat.it)
[3] *Italian National Institute of Statistics* (email: sabbi@istat.it)

The activities carried out for the update of ISTAT's Territorial Bases have led to the production of cartography derived from the integration of numerous geographically diverse data, especially at the thematic level. Therefore, the new statistical geography is highly valuable not only as a basis for disseminating census data or data resulting from surveys conducted by the Institute but also for spatial analyses concerning various social and territorial aspects. One of these aspects is the dynamics of inhabited localities. ISTAT inhabited localities have always been a cornerstone, primarily for the analysis of dynamics related to population and productive activities. In 2021, this geographic base, due to numerous innovations both in terms of graphics and information, allows for additional analyses related to various social aspects(transportation, infrastructure, leisure and recreational activities, etc.). In this paper, we will describe some simple examples of the dynamics of inhabited localities related to territorial phenomena (land consumption,) and those concerning population dynamics (depopulation, migration towards adjacent locations). As examples of significant changes in the surface area of localities, we can describe the cases of Udine and Sassari. The former, for a better resolution of the design, experiences a decrease in its surface by 21.5% compared to 2011. In contrast, Sassari increases its surface by just under 50% due to a merger with an adjacent locality. Moreover, by simple GIS algorithms for spatial analysis, is it possible the measurement of urban expansion in major population centers. From this, it can be deduced that among the main cities: - The inhabited center 'Milan' covers an area of approximately 75,098.1 hectares, expanding within 153 municipalities; - The urban area associated with Naples covers 49,957 hectares, involving the territory of 555 municipalities; - The urban area of Rome extends over approximately 51,072.7 hectares, expanding within 381 municipalities.

References

(1) Laaribi A., Peters L. 2019. GIS and the 2020 Census: Modernizing Official Statistics. Redlands. California: Esri Press.
(2) Lipizzi f. And mugnoli s. 2017. Profili e Dinamiche delle Località abitate in Italia. In Istat (eds). Forme, livelli e dinamiche dell'urbanizzazione in Italia. Istat, pp 39-58.
(3) UNITED NATIONS. 2021. Handbook on the Management of Population and Housing Censuses. Series F No. 83. New York: United Nations.

INFERENCE FOR BIG DATA ASSISTED BY SMALL AREA METHODS: AN APPLICATION ON SDGS SENSITIVITY OF ENTERPRISES IN ITALY

Monica Pratesi[1], Gaia Bertarelli[2]

[1] *Italian National Institute of Statistics* (email: monica.pratesi@unipi.it)
[2] *University of Venice "Ca'Foscari"* (email: gaia.bertarelli@unive.it)

In this study, we propose a new method to estimate the sustainable development goals (SDGs) sensitivity of enterprises in Italy at the provincial level using web-scraping data (a nonprobability sample) because this value is not surveyed by the Italian National Institute of Statistics. The proposed method uses a probability sample to reduce the selection bias of estimates obtained from the nonprobability sample in the context of small area estimation and integrates the nonprobability and probability samples using a double robust estimator that combines (i) propensity weighting to improve the representativeness of the nonprobability sample and (ii) a statistical model to predict the units that are not in the nonprobability sample. A bootstrap procedure for estimating variance is also proposed. To validate the proposed method, a Monte Carlo simulation, and an application with real data for e-commerce prevalence were performed. Results show that the proposed method allows the correction of bias from the nonprobability sample while maintaining a good level of estimate reliability.

BRINGING TOGETHER DIFFERENT DATA SOURCES IN ITALY: THE FOSSR PROJECT

Claudia Pennacchiotti[1], Gabriella D'Ambrosio[2], Primerano Ilaria[3]

[1] *CNR - Institute for Research on Population and Social Policies (CNRIRPPS)*
(email: claudia.pennacchiotti@cnr.it)
[2] *CNR Institute for Research on Population and Social Policies (CNR-IRPPS)*
(email: gabriella.dambrosio@cnr.it)
[3] *CNR Italy - Institute for Research on Population and Social Policies (CNR-IRPPS)*
(email: ilaria.primerano@cnr.it)

FOSSR project, funded in 2022 within the PNRR, aims to create a Research Infrastructure (RI) for social sciences, in compliance with Open Science and FAIR principles, to strengthen, among social sciences researchers, the awareness and knowledge on both data management and advanced statistical methods and techniques, via innovative interfaces (1). Moreover, aims to play a crucial role in addressing the key societal challenges, ensuring that national and European policy making is built on the highest-quality socio-economic data. Along the lines of the European Open Science Cloud project, FOSSR-RI represents a shared and simplified data access, in which innovative services for data collection, data curation and data analysis on economic and societal change are integrated. Among them, FOSSR RI builds up an Italian Life Course Observatory connecting data from already existing international surveys and infrastructures (Growing Up in Digital Europe - GUIDE; the Gender and Generations Survey - GGS; and the Survey of Health, Ageing and Retirement in Europe - SHARE-ERIC, all included in the ESFRI Roadmap 2021) with the first Italian Online Probability Panel - IOPP (in collaboration with ISTAT) which enriches data by integrating information on social and political attitudes and values of the Italian population (2). Bythis way, FOSSR RI makes available FAIR and updated meta-dataset covering the whole life cycle. Acknowledging the need to build new tools and foster existing infrastructures providing statistical information to support the future actions of researchers, institutions and policy makers, the paper aims to exploit the added value resulting from the integration of the above-mentioned surveys, enhancing synergies among them and providing scholars with a single access point to high quality and FAIR social science data in the perspective of a Life Course Observatory.

References

(1) Duşa A, Nelle D, Stock G, Wagner GG. Facing the Future: European Research Infrastructures for the Humanities and Social Sciences, eds. Berlin: Verlag; 2014.
(2) Callegaro M., Baker RP., Bethlehem J, Göritz AS., Krosnick JA, Lavrakas PJ. Online panel research: A data quality perspective, eds. New Jersey: John Wiley & Sons; 2014.

ASSESSING OPENNESS OF SOCIAL DATA PLATFORMS: A MIXED METHOD APPROACH

Paolo Landri[1], Luciana Taddei[2]

[1] *National Research Council of Italy, Institute for Research on Population and Social Policies (CNR-IRPPS)* (email: paolo.landri@cnr.it)
[2] *National Research Council of Italy, Institute for Research on Population and Social Policies (CNR-IRPPS)* (email: luciana.taddei@cnr.it)

Digital transformation fostered the proliferation of several processes of social change, both in individual behavior, in public and social life, and in scientific work. Within the Open Science scenario, the "platformization", i.e., the rapid proliferation of open, accessible, and easy-to-use data platforms, offers broad research insights, from theorizing theoretical models to defining innovative research methodologies. Indeed, the promotion of open data delivered by digital platforms poses new challenges and opportunities in social science research.The socio-technical approach identifies complex relationships among social and technical aspects (1), underlining barriers and possibly overcoming them. In particular, the notions of "black box", "boundary objects" and "standardized packages" help to analyze steps in building platforms (2). Furthermore, several studies have focused on the definition of the key indicators of openness able to characterize digital platforms (3). Moving from this framework, this contribution provides a critical review of how openness has been progressively measured with the twofold aim of highlighting the common pitfalls encountered by researchers while studying openness and defining an indicator of openness of social data platforms, able to integrate theoretical models and quantitative indexes following a mixed-method approach.

References

(1) Zuiderwijk, A., Janssen, M. (2014). Barriers and Development Directions for the Publication and Usage of Open Data: A Socio-Technical View. In: Gascó-Hernández, M. (eds) Open Government. Public Administration and Information Technology, vol 4. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-9563-5_8

(2) Landri, P. (2020). Educational leadership, management, and administration through actor-network theory. Routledge, London.
(3) Setzke, Böhm, and Krcmar (2019), Platform Openness: A Systematic Literature Review and Avenues for Future Research, 14th International Conference on Wirtschaftsinformatik, February 24- 27, 2019, Siegen, Germany.

DESIGNING THE ITALIAN ONLINE PROBABILITY PANEL: INNOVATIONS AND CHALLENGES TO FOSTEROPEN SCIENCE

Nicolò Marchesini[1], Luciana Taddei[2], Francesco Visconti[3]

[1] *National Research Council of Italy, Institute of Research on Population and Social Policies (CNR-IRPPS)* (email: nicolo.marchesini@cnr.it)
[2] *National Research Council of Italy, Institute of Research on Population and Social Policies (CNR-IRPPS)* (email: luciana.taddei@cnr.it)
[3] *National Research Council of Italy, Institute of Research on Population and Social Policies (CNR-IRPPS)* (email: francesco.visconti@cnr.it)

Probability-based online panels are a means to gather accurate and reliable data for research purposes. These panels have become popular because they provide researchers with access to a diverse and representative sample of the population, ensuring that the findings are both valid and generalizable (1). Italy lacks web-based, research-oriented (non-commercial) probabilistic panel surveys. This article provides an overview of the Italian Online Probability Panel (IOPP), a transformative tool developed within the Fostering Open Science in Social Science Research (FOSSR) project. The discussion encompasses the challenges encountered in realizing the IOPP, emphasizing its potential as a significant resource for the Italian social science research community, capable of driving positive interdisciplinary change and useful outcomes for institutions and policy makers (2). This is done through an in-depth description of the design of IOPP, the first online probability panel of Italian and foreign population in Italy. The article outlines the sampling design employed to recruit 10.000 panellists representative of the reference population and discusses the methodologies applied. Additionally, it details the content of the IOPP, covering various life stages and addressing a broad spectrum of topics, including families, housing, working life, income, vulnerability, gender, inequality, poverty, social and political attitudes, and values. The paper proceeds by outlining the structure of the IOPP, organised into five annual survey waves, featuring a core questionnaire and rotating modules. The open access policy is highlighted, promoting transparency and enabling researchers, policymakers, and the public to explore, analyse, and comprehend social dynamics. The contribution concludes by critically examining potential innovations and challenges associated with the design and development of this novel research tool, drawing insights from other European experiences (3). It explores how various stakeholders, including researchers, NGOs, advocacy groups, and the public, can actively engage with and leverage the open data provided by the IOPP.

References

(1) Callegaro M, Villar A, Krosnick J, Yeager D. A Critical Review of Studies Investigating the Quality of Data Obtained with Online Panels. 2014.
(2) Scherpenzeel A. Data collection in a probability-based internet panel: How the LISS panel was built and how it can be used. Bull Sociol Methodol. 2011;109(1):56-61.
(3) Arnesen S. A Guide to The 2017 European Internet Panel Study (EIPS). NORCE Norwegian Research Centre and University of Bergen, Bergen, Norway.

STATISTIC FOR ENVIRONMENT: TOURISM AND SUSTAINABILITY IN ITALY

Corrado Crocetta[1], Antonella Massari[2], Paola Perchinunno[3], Samuela L'Abbate[4]

[1] *University of Bari Aldo Moro* (email: corrado.crocetta@uniba.it)
[2] *University of Bari Aldo Moro* (email: antonella.massari@uniba.it)
[3] *University of Bari Aldo Moro* (email: paola.perchinunno@uniba.it)
[4] *University of Bari Aldo Moro* (email samuela.labbate@uniba.it)

The United Nations approved, in September 2015, the Sustainable Development Agenda and the related 17 goals to be achieved by 2030. Individual Italian regions are, also, called upon to contribute to the achievement of these goals. Sustainable tourism, defined as a form of tourism that respects the resources on which the very future of the sector depends, has been attributed the functions of renewing the cultural pride of the host communities, empowering local communities, and protecting biodiversity. The objective of this work is the analysis of the environmental impacts that the tourism sector produces are explored. Often the development of this sector is useful for facilitating growth in less developed areas, but the effects on the environment are not considered. The available data will be analyzed at a provincial level through multivariate statistical methodologies (Totally Fuzzy and Relative method) and density-based spatial clustering methods (DBScan), which allow identifying contiguous areas with high levels of sustainable tourism. These aspects should guide the way for the distribution of resources and investments, as currently, not all Italian regions start from the same conditions.

Young people's awareness and attitudes towards climate change: empirical evidence from southern Italy

Crescenza Calculli[1], Angela Maria D'Uggento[2], Alessio Pollice[3], Nunziata Ribecco[4]

[1] *University of Bari Aldo Moro* (email: crescenza.calculli@uniba.it)
[2] *University of Bari Aldo Moro* (email: angelamaria.duggento@uniba.it)
[3] *University of Bari Aldo Moro* (email: alessio.pollice@uniba.it)
[4] *University of Bari Aldo Moro* (email: nunziata.ribecco@uniba.it)

In Italy, as in the rest of the world, disgraceful events caused by climate change are occurring more frequently: air pollution, water bombs, overflowing rivers, storm surges eroding the Italian coasts. The effects not only disrupt the natural landscape, but also affect economic activities, destroying businesses and private property. Recently, Cop28 (1) highlighted the urgency of tackling climate change, which also threatens food security and health worldwide. In Italy, Ipsos (2) reports that extreme events, especially heat waves and high temperatures, are the biggest concern for 80% of respondents. Furthermore, 60% of Italians believe that missed opportunities in the past have prevented effective climate action and that it is too late for meaningful intervention in the current circumstances. This study, grounded in the conviction that the key to a brighter future lies in the active involvement of young individuals in addressing global climate policy issues, aims to assess the awareness and concerns related toclimate change among approximately 1,700 high school students in Apulia. The survey, which was conducted in 2019 as part of the Ministry of University Project for Scientific Degrees in Statistics through a web questionnaire, involves the application of machine learning techniques for data analysis (3). The findings underscore the heightened sensitivity of young individuals towards global environmental challenges and their proactive inclination to advocate for sustainable policies from policymakers. The youth express a belief in the transformative impact of individual actions on fostering a more sustainable future, identifying specific eco-friendly practices, green innovation and advancing education for sustainable development as key axes. Unlike their Italian adult counterparts, the youth maintain optimism, harboring hope that mitigating the planetary collapse is still feasible.

References

(1) Nevitt, M. (2023) Assessing COP28: The New Global Climate Deal in Dubai. Just Security, https://ssrn.com/abstract=4667941I

(2) Ipsos (2023). Global Predictions Survey 2024

(3) James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). Unsupervised Learning. In: An Introduction to Statistical Learning. Springer Texts in Statistics, vol 103. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-7138-7_10

Practical Consequences of Wolpert's Theorem in Addressing the Issue of Missing Environmental Data

Emanuele Barca

Water Research Institute, National Research Council (email: emanuele.barca@ba.irsa.cnr.it)

The analysis of extensive space-time climatic datasets holds immense potential for shedding light on critical environmental issues. However, the accessibility of such information is often hindered by the presence of missing data, posing a substantial threat to the reliability of conclusions drawn from space-time data analysis. Consequently, the right selection of an effective method for recovering missing data becomes crucial, given the real risk that a poor dataset reconstruction may distort statistical outcomes and mislead decision makers. The conventional approach of designating the state-of-the-art method as the optimal choice is rejected following Wolpert's theorem. This theorem asserts that no single method universally applies successfully to any dataset. Instead, our proposed methodology operates under the assumption that missing data within specified loss percentages minimally alters the shape and parameters of the original complete data distribution. Subsequently, a curated selection of data recovery methods is made, ensuring diversity in their theoretical basis to capture various features from the incomplete dataset. The goal is to gather insights from different perspectives. Ultimately, the reconstructed dataset exhibiting the highest similarities in terms of probability distribution with the incomplete dataset is deemed the most accurate, providing practitioners and researchers, who may not be experts in missing data issues, a robust strategy for incomplete dataset reconstruction. The actually applied methods are Expectation Maximization (EM) algorithm for multivariate Gaussian data, MICE (Multiple Imputation Chained Equations) based on the predictive mean matching method and the Cubist method. Time series are related to Faeto, Biccari, Orsara and Troia (located in the Southern Italy) precipitation stations of the Civil Protection Service Apulia Reg. Agency.

VOTING AS A SIGN OF ITALIAN YOUNG PEOPLE'S POLITICAL WILLPOWER

Luigi Fabbris[1], Angela Maria D'Uggento[2], Ilaria Pepe[3], Ruggiero Quarato[3]

[1] *University of Padova* (email: luigi.fabbris@unipd.it)
[2]*University of Bari Aldo Moro* (email: angelamaria.duggento@uniba.it)
[3] *University of Bari Aldo Moro*
[4] *University of Bari Aldo Moro* (email: r.quarto@geo.uniba.it)

Recent dramatic events such as Pandemic, wars and the catastrophic effects of climate change should promptus to think about a different relationship between people's quality of life and their environment. However, crisis situations should be an opportunity to look for new solutions, create incentives for innovation and develop a long-term strategy for collective well-being. Today's society is characterized by unsustainable consumption patterns that are the result of excessive individualism (1). It is necessary to reverse this trend by identifying the current critical behaviors in order to develop future paths that are at the same time flexible, positive and desirable, knowing that individual and collective commitment can influence the course of events. Based on these assumptions, this study aims to get to know the world of associationism better, considered as a possible means of spreading positive principles of solidarity among young people in order to build a better society (2). Around 500 people were interviewed using a web questionnaire, with participants almost evenly divided between members with active participation in voluntary organizations and people who had never had such an experience. By comparing these two groups, it was possible to understand the key characteristics of those who volunteer and what drives a person to adhere to such principles and become a better person, family member, friend and citizen as a result. From the multivariate analysis, some interesting results have emerged that are useful to promote and improve the model of volunteering as a form of solidarity and expression of social capital, a driving force for a proactive attitude towards the future. Forms of social association such as volunteering could facilitate the overcoming of difficulties related to an uncertain future and the search for work. The emphasis on the possibility of influencing social systems urges a positive vision of the future that allows for the existence of plausible, possible, probable and preferred futures (3,4).

References

(1) Bell W., Foundations of Futures Studies: Human Science for a New Era, Transation Publishers, New Brunswick, London, 2003.
(2) Comper, C. (2022) Visioni del futuro nelle organizzazioni di volontariato tra futuri desiderati e futuri possibili. Italian Institute for the Future, Napoli. www.instituteforthefuture.it.
(3) Pacinelli, A. Pacinelli A., I metodi della previsione, in Arnaldi S., Poli R. (a cura di), La previsione sociale.Introduzione allo studio dei futuri, Carocci, Roma, 2012.
(4) Poli R., Lavorare con il futuro, Egea, Milano, 2019.

MENTAL HEALTH MATTERS: A STUDY OF ACADEMIC WELL-BEING

Lucia Di Stefano[1], Paolo Parra Saiani[2] , Enrico Ivaldi[3]

[1] *University of Genoa* (email: lucia.distefano@edu.unige.it)
[2]*University of Genoa* (email: paolo.parra.saiani@unige.it)
[3] *IULM* (email: enrico.ivaldi@iulm.it)

The Covid-19 pandemic has turned into a global crisis that has thrown the world into uncertainty and upheaval, disrupting daily life and obscuring future plans. Its ramifications have presented in various spheres, from health to social, economic and political, affecting quality of life and well-being. Accordingly, this project seeks to explore the mental health status of Italian academic social scientists in the context of the challenges posed by the pandemic. Given their central role in societal decision-making as well, studying their experiences during this period is crucial. The pandemic has exacerbated existing mental health problems among academics, manifesting in higher levels of depression, stress, anxiety, and financial strain, as well as disruptions in social networks and increased burnout. The shift to virtual learning and the disruption of social connections have not only affected students but have also had a profound impact on the work of professors and researchers, with women professors having a disproportionate burden of responsibility. Moreover, the relentless pursuit of academic duties amid the chaos of the pandemic has further challenged individuals, necessitating the recognition of personal limits and the setting of boundaries to achieve balance. To investigate these challenges, a web survey was conducted among the active community of social scientists in Italian universities. The survey, distributed via institutional email addresses, aimed for a representative sample in terms of gender, sector, and geographical distribution. The questionnaire, divided into three sections, explored various aspects of respondents' mental and physical well-being, work-related stress, and the impact of Covid-19 on their lives. The analysis of collected data involved two phases: an Exploratory Factor Analysis (EFA) to identify latent factors within the questionnaire sections and a Logistic Regression to assess the relative risk on respondents' health perception. The findings underscore the multifaceted nature of health perceptions, ranging from individual well-being to the societal and environmental contexts in which illness occurs. The academic landscape has undergone significant transformations, marked by standardized teaching practices and heightened pressure for publication, leading to discomfort and stress among scholars. Amidst these challenges, understanding and addressing the mental health needs of academics emerge as imperative tasks in navigating the post-pandemic academic realm.

STRUCTURAL VULNERABILITY AND ALCOHOL CONSUMPTION: PASTOS INDIGENOUS PEOPLE IN SOUTH-WESTERN COLOMBIA

Diego Ignacio Meza Gavilanes

Pontifical Gregorian University (email: d.meza@unigre.it)

Alcohol consumption as a social phenomenon and as a public health problem defies the quality of life of many populations, particularly those where poverty and inequalities are highest. In this sense, the concept of structural vulnerability has been used to explain the causes of health inequalities and drug, alcohol, and tobacco consumption in marginalised communities (1,2,3). Structural vulnerability captures the missing link between clinical medicine and social science to explain how social, economic, and political hierarchies produce and shape negative health outcomes. Therefore, based on survey research conducted in the Gran Cumbal indigenous reservation in southwestern Colombia (N=979), a logistic regression model shows how people exposed to structural vulnerability are more prone to become drunkards. Indigenous people are structurally vulnerable (age, housing, occupation, education, life satisfaction) when these factors interfere with their ability to access or benefit from a life wellbeing which degenerates into habitual alcohol consumption.

References

(1) Bourgois, P., Holmes, S.M., Sue, K. and Quesada, J. (2017). Structural vulnerability: operationalizing the concept to address health disparities in clinical care. Academic medicine: journal of the Association of American Medical Colleges, 92(3), pp. 299–307.
(2) Herrick, C. (2017). Structural violence, capabilities and the experiential politics of alcohol regulation. In C. Herrick and D. Reubi (Eds), Global Health and Geographical Imaginaries, London: Routledge.
(3) Quesada, J., Kain, L. and Bourgois, P. (2011). Structural Vulnerability and Health: Latino Migrant Laborers In the United States. Medical Anthropology, 30(4), pp.339–362.

CLUSTERING AND MODEL-BASED COMPOSITE INDICATORS FOR ENVIRONMENTAL ANALYSIS

Maurizio Vichi

University of Rome "La Sapienza" (email: maurizio.vichi@uniroma1.it)

The objective of this presentation is to explore various new methods applicable to multivariate statistical analysis of environmental data. The overarching goal is to determine an appropriate number of clusters for a multivariate set of units, computing a model-based composite indicator, and assuming the presence of a ranking among clusters and hence, considering the ordered arrangement of centroids. Given our aim to establish a detailed ranking among units, the desired number of clusters should be maximized, ensuring well-separated centroids.

References

(1) Mariaelena Bottazzi Schenone, Elena Grimaccia, Maurizio Vichi (2024). Structural equation models for simultaneous modeling of air pollutants, Environmetrics, https://doi.org/10.1002/env.2837.
(2) Mariaelena Bottazzi Schenone, Maurizio Vichi (2024). Clustering for ranking multivariate data by Linear Ordered Partitions, submitted.
(3) Mariaelena Bottazzi Schenone, Elena Grimaccia, Maurizio Vichi (2024). Assessing environmental quality by clustering a structural equation model based index: An application to European cities air pollution, submitted.

FUNCTIONAL DATA ANALYSIS AND GROUP LASSO INTEGRATION FOR ASSESSING CHEMICAL AND METEOROLOGICAL INFLUENCES ON PM10 CONCENTRATION

Tonio Di Battista[1], Adelia Evangelista[2], Annalina Sarra[3], Christian Acal[4], Ana M. Aguilera[5], Sergio Palermi[6]

[1] *University G. d'Annunzio* (email: dibattis@unich.it)
[2] *University G. d'Annunzio* (email: adelia.evangelista@unich.it)
[3] *University G. d'Annunzio* (email: annalina.sarra@unich.it)
[4] *University of Granada* (email: chracal@ugr.es)
[5] *University of Granada* (email: aaguiler@ugr.es)
[6] *Agency of Environmental Protection of Abruzzo* (email: s.palermi@artaabruzzo.it)

Particulate matter (PM) is a dangerous airborne pollutant with harmful impacts on human health. Legislative bodies pay particular attention to ambient breathable particles with a diameter of less than 10 micrometers (PM 10). In urban areas PM 10 levels result from a combination of elements, including regional background concentrations, urban emissions, and traffic-related sources. Additionally, meteorological conditions exert a crucial influence. This work is aimed at analysing and identifying the significant effects of chemical and local meteorological variables on the evolution of PM 10 concentration in the Abruzzo region (Italy), by adopting a FDA approach. A comprehensive three-step FDA methodology is introduced to estimate a functional response variable considering multiple functional covariates. The process initiates with the estimation of univariate Functional Principal Component Analysis (FPCA) for each pertinent functional variable. Following this, we define a Multiple Functional-Functional Linear Regression (MFFLR) model and estimate it through multiple linear regression (1). This entails establishing connections between the functional response variable and its Principal Components (PCs), specifically those significantly correlated with the predictor variables PCs (2). To ensure precise model selection, we employ group Lasso estimation (3) for each multiple linear model associated with the response PCs. A considerable portion of the variability in PM 10 can be explained by the first two Principal Components (PCs). Using the group Lasso criterion, we identified the PCs of functional predictors that significantly affect the first two PCs of the target variable. Key meteorological variables, such as pressure, rain, relative humidity, temperature, solar radiation, and nitrogen dioxide, merged as influential factors contributing to PM10 concentration accumulation. The proposed approach, integrating FDA and group Lasso, proves promising for addressing challenges in in the context of MFFLR. It offers a robust and interpretable model applicable to environmental research.

References

(1) C. Acal, M. Escabias,A.M. Aguilera, M.J. Valderrama. COVID-19 Data imputation by Multiple Function-on-Function Principal Component Regression. Mathematics, 9: 1237, 2021.
(2) A.M. Aguilera, F.A. Ocaña, M.J. Valderrama. Forecasting time series by functional PCA. Discussion of several weighted approaches. Computational Statistics, 14:443–467, 1999
(3) M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B, 68(1): 49–67, 2006.

GENERALIZED REGULARIZED REDUCED-RANK REGRESSION MODELS WITH MIXED RESPONSES AND MIXED PREDICTORS

Lorenza Cotugno[1], Mark De Rooij[2], Roberta Siciliano[3]

[1] *University of Naples Federico II* (email: lorenza.cotugno@unina.it)
[2] *Leiden University* (email: rooijm@fsw.leidenuniv.nl)
[3] *University of Naples Federico II* (email: roberta.siciliano@unina.it)

This paper proposes a Generalized Mixed Regularized Reduced Rank Regression model (GMR4) for mixed response and predictor variables. An algorithm will be developed and implemented in R, tested using simulation studies, and applied to an empirical data set. We will use data from the European Commission's Eurobarometer Surveys (Jan-Feb 2023), with a specific focus on residents of the Netherlands. Reduced Rank Regression models are regression models for multiple outcome variables. These models over time have been used for different types of response variables: numeric (2), binary (1), and ordinal response variables. To deal with different types of predictor variables, Optimal Scaling will be used (3). Furthermore, this model includes a penalty on the coefficients to address challenges associated with high-dimensional data, such as having an excessive number of predictors compared to observations and the presence of multicollinearity among the predictors. Penalties can be applied using Ridge, Lasso, or Elastic Net. For the estimation of the parameters, the Majorization Minimization algorithm will be presented. Furthermore, a small Monte Carlo simulation study is set up to investigate how well the algorithm retrieves population parameter value.

References

(1) De Rooij, Mark. A new algorithm and a discussion about visualization for logistic reduced rank regression. Behaviormetrika (2023): 1-22.
(2) Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. Journal of multivariate analysis, 5(2):248–264.
(3) Meulman, J., Anita J. van der Kooij, and Kevin LW Duisters. ROS regression: Integrating regularization with optimal scaling regression. (2019): 361-390.

EXAMINING SPARSE ARCHAEOLOGICAL DATA: ADVANTAGES AND DRAWBACKS OF SIMPLE CORRESPONDENCE ANALYSIS AND ITS VARIANTS

Rosaria Lombardo[1], Eric J. Beh[2]

[1] *University of Campania "Luigi Vanvitelli"* (email: rosaria.lombardo@unicampania.it)
[2] *University of Wollongong, Stellenbosch University* (email: ericb@uow.edu.au)

This study delves into the analysis of a substantial and sparsely populated archaeological dataset, characterized by overdispersion in the cell frequencies. Obtained from an archaeological site at Capo Milazzese, Sicily, the dataset is summarised in a 20 × 13 contingency table, cross-classifying 374 artifacts distributed among 13 huts (1). With correspondence analysis recognized as a prominent technique for visualizing the nature of the association between categorical variables (1, 2), our primary aim is to employ variants based on the Cressie-Read family of divergence statistics (3). This specialized approach seeks to reveal associations within the distinctive context of the Capo Milazzese dataset. We address the challenge of overdispersion by investigating how these divergence statistics capture nuanced patterns in large, sparse archaeological data.

Our analysis utilizes Pearson's chi-squared statistic, the Freeman-Tukey statistic, the likelihood ratio statistic, and modified versions of some of these that have been extensively discussed in the literature. The unique structure of our dataset calls for a tailored examination of correspondence analysis, with a specific focus on evaluating the distribution of residuals for these divergence statistics.

Preliminary findings highlight the distribution of the residuals obtained from each of the divergence statistics considered within the Cressie-Read family, providing insights into their applicability for large, sparse, and overdispersed archaeological datasets.

This study contributes valuable insights into the utilization of variants of correspondence analysis within the Cressie-Read family for understanding associations in challenging sparse datasets. By concentrating on the case study involving the data from the Capo Milazzese, Sicily, site our findings present a framework for researchers working with similar large, sparse, and overdispersed datasets. The study underscores the significance of tailored statistical approaches for extracting meaningful insights in such contexts.

References

(1) Alberti, G. (2017). New light on old data: Toward understanding settlement and social organization in Middle Bronze Age Aeolian Islands (Sicily) through quantitative and multivariate analysis. Journal of Archaeological Science: Reports, 11, 310 – 329.
(2) Beh, E. J. and Lombardo, R. (2014). Correspondence Analysis: Theory, Practice and New Strategies. Wiley: Chichester.
(3) Beh, E. J. and Lombardo, R. (2024). Correspondence analysis using the Cressie-Read family of divergence statistics. International Statistical Review doi: 10.1111/insr.12541.

USING SUPERVISED ANN TO INPUT MISSING CATEGORICAL DATA

Francesco D. D'Ovidio[1], Angela M. D'Uggento[2], Najada Firza[3], Alessandro Pagano[4], Ernesto Toma[5]

[1] *University of Bari Aldo Moro* (email: francescodomenico.dovidio@uniba.it)
[2] *University of Bari Aldo Moro* (email: angelamaria.duggento@uniba.it)
[3] *University of Bari Aldo Moro* (email: najada.firza@uniba.it)
[4] *University of Bari Aldo Moro* (email: alessandro.pagano@uniba.it)
[5] *University of Bari Aldo Moro* (email: ernesto.toma@uniba.it)

Investigations, surveys and other research studies frequently encounter missing data. Regardless of the method used to select individuals for a survey or study, there may be unit non-response, where some individuals decline to participate or the researcher is unable to contact them. Furthermore, among those who participate, some may drop out of the study early, resulting in loss to follow-up. Additionally, some individuals may leave certain survey items unanswered, whether due to refusal or lack of knowledge; this is known as item non-response.
Similarly, in investigations involving physical measurements, some individuals may not be measured or their measurements may fail or be missing.
This paper presents a methodological framework for developing an automated data imputation model based on Artificial Neural Networks. The model was designed for recursive supervision with the aim of developing an appropriate algorithm for this type of supervision. The study analysed a dataset of over 4,600 cases, which had thousands of item non-responses to a key question that required ordinal-categorical answers, and, consequently, methods that were suitable for continuous or discrete values were not applicable.
Several architectures and learning algorithms were tested for the multilayer perceptron to find the best imputation on the training and test sets due to the flexibility of ANN techniques. The answers in the test set were known but hidden. The artificial neural network (ANN) rules were applied to cases with true missing values (holdout set) only if their associated pseudo-probabilities were very high, while cases with lower pseudo-probabilities were recursively reprocessed until they reached the goal.
The results suggest that this approach significantly improves the quality of a database with missing values in data sets that contain categorical variables. It may provide valuable insights into sensitive areas such as undisclosed incomes.

A BAYESIAN QUANTILE REGRESSION MODEL IN THE ITALIAN JUDICIARY FRAMEWORK

Carlo Cusatelli[1], Massimiliano Giacalone[2], Eugenia Nissi[3]

[1] *University of Bari Aldo Moro* (email: carlo.cusatelli@uniba.it)
[2] *University of Naples Federico II* (email: massimiliano.giacalone@unina.it)
[3] *University of Chieti-Pescara Gabriele D'Annunzio* (email: nissi@unich.it)

The efficiency of the Italian judicial system is a topic of great relevance for the society and the economy of the country. The Italian judicial system presents some structural and organizational problems that slow down its functioning and compromise the quality of justice. Among these problems, we can mention the scarcity of human and material resources, the complexity of the rules and procedures, the length of the trials and the overcrowding of the prisons. To improve the efficiency of the Italian judicial system, legislative, administrative and cultural interventions are needed that favor simplification, digitalization, specialization and accountability of the judicial operators. Using data disaggregated at the district level, we aim to measure and compare the efficiency of Italian judicial offices.

To this end, quantile regression is a useful tool to model the conditional distribution of a response variable given some covariates. However, traditional quantile regression methods may not be robust to outliers or heavy-tailed data. In this paper, we propose a Bayesian quantile regression model based on the Skew Exponential Power (SEP) distribution, which can account for different levels of tail decay and asymmetry. We apply our model to the Italian judiciary framework, using a hierarchical structure to capture the heterogeneity and correlation among the courts, and we employ a dynamic model averaging approach to select the relevant covariates, finding that the SEP distribution provides a good fit to the data and reveals some interesting patterns in the conditional quantiles of the response variable. Our model can be useful for policy makers and practitioners who want to monitor and improve the efficiency of the judicial system.

THE USE OF MACHINE LEARNING TECHNIQUES IN SOCIAL STATISTICS: THE HEALTHCARE CONTEXT

Margaret Antonicelli[1], Filomena Maggino[2], Sofia Urbani[3]

[1] *La Sapienza, University of Rome* (email: margaret.antonicelli@uniroma1.it)
[2] *La Sapienza, University of Rome* (email: filomena.maggino@uniroma1.it)
[3] *La Sapienza, University of Rome* (email: sofia.urbani@uniroma1.it)

Artificial intelligence is revolutionizing the world in many economic sectors, and healthcare is no exception. Thanks to its innovative way of selecting information, analysing large quantities of data and learning from it with machine learning, artificial intelligence is becoming an increasingly important tool for improving the healthcare system. Above all, the use of machine learning techniques in healthcare can help to improve patient management, reducing waiting times and ensuring that healthcare resources are used more efficiently. For example, the use of these techniques can be employed to predict patient flow in a hospital, allowing staff to organize appointments and resources more effectively. This can lead to a better experience for patients and less overhead for doctors and nurses. As a direct consequence of greater efficiency, in the management of both patients and medical personnel, equipment and supplies, we find a clear drop in healthcare costs, with the possibility of using funds for improving many aspects such as research efficiency of machinery and structures hospital.

This study, through a web scraping process carried out on the websites of all Italian Local Health Authorities, investigates the demographic and economic determinants of digitalisation. To conduct the analysis, a composite indicator was first developed and calculated which quantified the level of digitalisation of the Local Health Authorities and, subsequently, the information obtained was performed within a Tobit model. This allowed us to highlight the critical issues present within the individual units, allowing us to formulate critical reflections regarding the digitization policies adopted.

This study offers important theoretical and practical implications and enriches the current literature on the topic of digitalization in healthcare.

HOUSEHOLD ECONOMICS AND DEMOGRAPHIC CHARACTERISTICS: THE CASE OF THE REAL ESTATE MARKET IN THE CITY OF BARI

Caterina Marini[1], Vittorio Nicolardi[2]

[1] *University of Bari Aldo Moro* (email: caterina.marini@uniba.it)
[2] *University of Bari Aldo Moro* (email: vittorio.nicolardi@uniba.it

Data that derive from variegated and autonomous sources of information are object of well-established fields of research that focus on analytical techniques to depict the most various socio-economic phenomena. The main issue, central in the scientific international debate, is the utilisation of the administrative data, as yielded by private organisations/enterprises/companies and public administrations, alongside the official statistics, as yielded by the National Institutes of Statistics. In theory, there is a multitude of information, very detailed and valuable, but its alignment is impeded by many problematic issues that involve the quality and variety of data, the juridical protection of privacy and the magnitude of database allocated in the data science scenario in most of the cases. In few words, the National Institutes of Statistics need flows of data that derive from different data collectionsfor administrative purposes, but the resolution of the involved issues isstill part of the scientific debate. In our recent works (1, 2), we succeeded in aligning administrative data and official statistics referring to the real estate phenomenon, from both administrative and market side, and we produced an algorithm to create the first Full Information Harmonised Real Estate Database (FIHRE-DB) limited to the city of Bari, anticipating the Italian National Institute of Statistics (ISTAT). The great achievement we obtained is showing the potentiality of FIHRE-DB to develop many spatial analyses by merging many other information, differently collected, with the census sections as defined by ISTAT. In this work, we show a practical application of that potentiality, and we develop an analysis that involved the household wealth and the distribution of household demographic characteristics based on the real estate market in the city of Bari. The assumptions we made and the outcomes we obtained are a proof of the validity of an analysis that the alignment of information can achieve.

References

(1) Marini C., Nicolardi V. Big data and Economic Analysis: The Challenge of a Harmonised Database. In Mariani P. and Zenga M., editors. Data Science and Social Research – II DSSR 2019. Studies in Classification, Data Analysis, and Knowledge Organization: 235-246, Springer Nature, 2019. DOI: 10.1007/978-3-030- 51222-4_18.
(2) Marini C., Nicolardi V. Administrative database and official statistics: The case of the real estate analysis. Italian Journal of Applied Statistics, 2021; 33: 83-95. DOI: 10.26398/IJAS.0033-004

TALL: A NEW SHINY APP OF TEXT ANALYSIS FOR ALL

Massimo Aria[1], Corrado Cuccurullo[2], Luca D'Aniello[3], Michelangelo Misuraca[4], Maria Spano[5]

[1] *University of Naples Federico II* (email:massimo.aria@unina.it)
[2] *University of Naples Federico II* (email:corrado.cuccurullo@unina.it)
[3] *University of Naples Federico II* (email:luca.daniello@unina.it)
[4] *University of Calabria* (email:michelangelo.misuraca@unical.it)
[5] *University of Naples Federico II* (email:maria.spano@unina.it)

In the era of big data, researchers across various disciplines face the challenge of analyzing extensive textual data spanning research articles, social media posts, customer reviews, and survey responses. These new sources harbor valuable insights applicable to advancing knowledge in several fields ranging from the social sciences to healthcare. Researchers aim to identify patterns, recognize trends, and extract meaningful information from textual data, employing advanced natural language processing (NLP) techniques and machine learning algorithms for tasks such as topic detection, polarity detection, and text summarization.
Moreover, the rise of digital platforms and the proliferation of online content have generated vast amounts of previously inaccessible textual data. Researchers tap into these resources to explore new research questions, validate existing theories, and develop novel insights. Computational tools facilitate the efficient processing and analysis of large text volumes, significantly reducing the time and effort required compared to manual methods. However, many researchers lack the necessary programming skills for effective textual data analysis, creating a demand for user-friendly text analysis tools. Despite the powerful capabilities of R and Python, acquiring proficiency in these programming languages often requires additional time or resources. This paper presents the first version of TALL - Text Analysis for All - a new R Shiny app that combines all the major text analysis advancements developed in recent years. TALL serves as a practical solution for researchers without programming skills, offering an intuitive interface that enables interaction with data and the execution of analyses without extensive programming knowledge. TALL provides a comprehensive workflow for data cleaning, pre-processing, statistical analysis, and visualization of textual data by combining state-of-the-art text analysis techniques into an R Shiny app.

INVESTIGATING TOPIC INTERPRETABILITY OF LEGAL CORPORA: A FUZZY TOPIC MODELLING APPROACH

Antonio Calcagnı[1], Arjuna Tuzzi[2]

[1] *University of Padova* (email: antonio.calcagni@unipd.it)
[2] *University of Padova* (email: arjuna.tuzzi@unipd.it)

Budget laws act as foundational frameworks governing a country's finances, intricately dictating fund acquisition and allocation. Interconnected with broader economic and administrative laws, they shape fiscal policies and guide government spending. However, their conventional written format often fails to adequately recognize expenditure chapters, limiting understanding of governmental actions. To provide an unsupervised way to aggregate chapters dealing with the same expenditure subject, we investigate the use of a fuzzy topic model for short corpora. In particular, we investigate the coherence and exclusivity of identified topics, aiming to remove incoherent or unimportant background topics that may inaccurately represent the short corpora. The adopted solution integrates local and global term weighting alongside dimensionality reduction techniques (e.g., SVD) to alleviate sparsity in word tokens and corpora features. In addition, it uses a fuzzy k-medoids algorithm which enables the soft clustering of words and documents into coherent topics. Finally, a case study based on the Italian budget law is conducted to evaluate the interpretability and meaningfulness of the expenditure chapters derived from the fuzzy topic modelling approach.

DYNAMIC COMMUNITY DETECTION FOR FRAMING ANALYSIS: CAPTURING FRAMES OVER TIME

Manuel Jesus Cobo Martin[1], Alberto Maria De Mascellis[2], Michelangelo Misuraca[3], Germana Scepi[4], Maria Spano[5]

[1] *University of Granada* (email: mjcobo@ugr.es)
[2] *University of Naples Federico II* (email: albertomaria.demascellis@unina.it)
[3] *University of Calabria* (email: michelangelo.misuraca@unical.it)
[4] *University of Naples Federico II* (email: germana.scepi@unina.it)
[5] *University of Naples Federico II* (email: maria.spano@unina.it)

In public opinion studies there are several theories about the so-called "frames", portions of information that individuals use to comprehend the world and that can be used to influence governments into arbitrary agendas. Framing Analysis is a paradigm aimed at studying the way these frames are created and how they interact with public opinion through mass media, to promote a particular definition of a problem, a causal interpretation, a moral evaluation and/or a recommendation for the object encased in such structures (1). Since in the contemporary big data society there is an overgrowth of this kind of documents, this paper aims to take further steps towards the development of an automatic processing technique for large quantities of news information, keeping in mind the "frame" definition in qualitative and quantitative literature. We rely on the well-established synergy between Framing Analysis and Community Detection (2), but with a diachronic perspective via Dynamic Community Detection techniques (3), capturing "snapshots" of newspaper information and mapping their evolution over the years.

We tested this approach on a case study phenomenon (the so-called Reddito di Cittadinanza), analysing texts from five newspapers of national relevance from 2018 to 2023, identifying not only the communities/frames that newsmakers have created to draw public's attention, but also how they increase (or decrease) in terms of concepts framed and how they change obeying the turns of the political, economic and social macro-events, based on the knowledge domain acquired from the debate on the case study. As a preliminary result of this strategy, we observed the presence of some frames and sub-frames (like cultural/artistic or judicial/public) that evolved over the years with new terms and information chunks, acquiring new meanings or even unprecedented turns.

References

(1) Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. Journal of communication, 43(4), 51-58.
(2) Fortunato, S., & Castellano, C. (2007). Community structure in graphs. arXiv preprint arXiv:0712.2716.
(3) Sarmento, R. P., Lemos, L., Cordeiro, M., Rossetti, G., & Cardoso, D. (2019). DynComm R Package—Dynamic Community Detection for Evolving Networks. arXiv preprint arXiv:1905.01498.

ECO-CENTRIC LENS: UNVEILING TOPICS IN SUSTAINABLE TOURISM FOR A GREENER FUTURE

Emma Zavarrone

University IULM (email: emma.zavarrone@iulm.it)

This study explores sustainable tourism using natural language processing (NLP) as its framework. The study utilises modern NLP approaches to detect topics and explore the complex nature of sustainability in tourism across the Asian area, with a specific focus on best practices. The study begins by providing a backdrop for sustainable tourism in Asia and emphasises the crucial role of NLP in understanding and analysing various practices. The research utilises knowledge from sustainable tourism and NLP literature to create a framework for further investigation into topic detection approaches.

An extensive examination is conducted, assessing different Natural Language Processing (NLP) methodologies, including conventional algorithms like Latent Dirichlet Allocation (LDA) and state-of-the-art transformer-based structures designed for NLP. The focus is on how effective they are in revealing hidden subjects and trends in large datasets that cover sustainable tourism practices in various Asian destinations.

The research explores a detailed comparison of these techniques, using case studies that reflect different Asian locations. This study examines different NLP techniques to understand their strengths and limits in capturing the many aspects of sustainable tourism. Furthermore, it examines how these approaches adjust to the distinct cultural, economic, and environmental aspects that define Asian destinations. The findings enhance our comprehension of sustainable tourism practices in Asia, highlighting the significance of NLP approaches in revealing valuable insights. The findings have significant significance for stakeholders, policymakers, and researchers, providing valuable recommendations for promoting the development of sustainable practices in the region.

Overall, this study, carried out from a "Eco-centric Lens," not only demonstrates the effectiveness of NLP techniques in uncovering subjects related to sustainable tourism but also emphasises the importance of Asian best practices. The study utilises modern NLP techniques to propose a path towards a more environmentally friendly and sustainable future for tourism, based on the diverse range of Asian sustainability projects.

Eco-centric Lens: Unveiling Topics in Sustainable Tourism for a Greener Future

Emma Zavarrone

University of Milan IULM (email: emma.zavarrone@iulm.it)

EXPLORING THE NARRATIVES ON ARTIFICIAL INTELLIGENCE IN THE CONTEXT OF THE 2030 AGENDA: A SOCIALMEDIA CONTENT ANALYSIS

Noemi Crescentini[1], Cristiano Felaco[2]

[1] *University of Naples Federico II* (email:noemi.crescentini@unina.it)
[2] *University of Naples Federico II* (email:cristiano.felaco@unina.it)

This contribution aims to reconstruct narratives surrounding the role of Artificial Intelligence in achieving the Sustainable Development Goals (SDGs) outlined in the 2030 Agenda. The 2030 Agenda, adopted at the international level by the United Nations aims to promote sustainable development on a global scale, addressing every region and nation on the planet. The 17 Sustainable Development Goals (SDGs) outline a multidimensional framework embracing different spheres of socio-economic and environmental development.
The realisation of some of these objectives can be facilitated using Artificial Intelligence (AI), which has the potential to improve the efficiency, equity and impact of actions taken. However, it is essential to note that the application of AI can also have negative implications, especially in the absence of an appropriate governance framework for managing of these technologies. In this context, there is a need for reflection on effective strategies to mitigate any risks arising from the adoption of AI in the pursuit of SDGs.
We performed a content analysis of social media posts in Italy to delve into these narratives, aiming to understand how social media platforms contribute to shaping public perceptions not only of the 2030 Agenda but also of the link between artificial intelligence and sustainable development.
The results will provide a detailed depiction of the prevailing narrative on how artificial intelligence aligns the 2030 Agenda, identifying key areas, the most discussed issues, prevailing tones, and ways in which the online community is engaged.

A NEW REVIVAL OF CONTENT ANALYSIS? USES, PURPOSES, AND CATEGORISATION SYSTEM IN THE ERA OF DIGITAL TRACES

Enrica Amaturo[1], Gabriella Punziano[2], Giuseppe Michele Padricelli[3]

[1] *University of Naples Federico II* (email: enrica.amaturo@unina.it)
[2] *University of Naples Federico II* (email: gabriella.punziano@unina.it)
[3] *University of Naples Federico II* (email: giuseppemichele.padricelli@unina.it)

Content analysis has witnessed a resurgence in the digital realm, adapting its established methodological principles to new forms of online content.
A significant portion of the data generated online, which previously was almost always user generated, now also results from algorithmic intervention. Therefore, researchers face new challenges related to the kind of objects of analysis elected as data, including their accessibility, availability, collection and analysis, as well as determining plausible research questions and valid cognitive interpretation.
Digital traces and algorithmic productions that include structured and unstructured semantics are the result of technical and cultural processes of co-construction. Distinguishing between user-created traces and those of "algorithmic media" is a crucial challenge, as it reveals connections between human and nonhuman actors and imposes new lenses on the researcher in order to understand their essence.
Content analysis, therefore, has become an essential approach to exploring Internet-related phenomena characterised by digital human and nonhuman traces. This work aims to promote methodological reflection by addressing cognitive inquiries and data accessibility in customising user sense production. It questions whether content analysis needs to adapt its categories to accommodate human-algorithm-driven information production and explores the concept of a new digital-hybrid content analysis perspective.

TALES OF FUTURE. A METHODOLOGICAL PROPOSAL FOR STUDYING IMAGINARIES IN DIGITAL AND TECHNOLOGICAL TRANSFORMATION

Suania Acampa

Southern Centre for Digital Transformation, University of Naples Federico II (email: suania.acampa@unina.it)

In the realm of social sciences, there has been a growing interest in examining how social actors perceive and anticipate technological innovations. Jasanoff and Kim (1) introduce the concept of "sociotechnical imaginaries," referring to the narratives, visions, and expectations about technology's future role and impact that are collectively held by a society. This ability of social actors to envision the technological future is not just theoretical but performative too: it actively shapes activities, investments, government policies, and legislative frameworks (2). Recent studies have particularly focused on the role of narratives to analyse this projective ability (3): they play a crucial role in identifying and elucidating how these imaginaries emerge as significant narratives about the sociotechnical future. This highlights the need for an analytical approach that leverages discursive practices to explore expectations related to emerging technologies and digital transformation processes. The research presented here aims to address this gap. It outlines a methodological approach based on exploratory sequential mixed methods digital design divided into different phases through which it will be possible to investigate the expectations, visions, and orientations that drive these narratives, as well as the emerging transformation processes and the themes they utilize. Additionally, it will reconstruct the network of actors involved in narrating the mission of digital and technological transformation. It begins with an analysis of official documents from European countries concerning emerging technologies, to arrive to constructing an opinion dictionary using a hybrid Opinion Mining approach, which combines supervised and unsupervised methods. This proposal can be a replicable tool for scholars and analysts in the field of innovation. Through this methodology, the study aims to semi-automatically reconstruct the narratives that underpin the imaginaries of the technological future, thereby gaining insights into how current expectations are already shaping ongoing change processes.

References

(1) Jasanoff S, Kim SH (2015). Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power. Chicago: University of Chicago Press.
(2) Borup, M., Brown, N., Konrad, K., & Van Lente, H. (2006). The sociology of expectations in science and technology. Technology analysis & strategic management, 18,3-4: 285-298.
(3) Bazzani, G. (2023). Futures in Action: Expectations, Imaginaries and Narratives of the Future. Sociology,57(2), 382–397.

DATA ANALYSIS APPLIED TO AN INNOVATIVE AND IMMERSIVE E-COMMERCE PLATFORM FOR THE PROMOTION OF MADE IN ITALY

Battista Masellis[1], Sergio Vitullo[2], Michele Di Lecce[3], Angelo Lamacchia[4], Giulio Setzu[5], Antonio Ruoto[6], Marco Calciano[7], Gianfranco Piscopo[8]

[1] *B.M.*
[2] Informatica Srl,
[3] *University of Naples Federico II*
[4] *University of Naples Federico II*
[5] *University of Naples Federico II*
[6] *University of Naples Federico II*
[7] *Zio Startup Srls*
[8] *University of Naples Federico II*

The virtual tour to a location where editorial works are conceived, designed and implemented becomes a totally innovative and interactive dimension of e-commerce, where it is possible to be guided through a multilingual shopping experience that is both a journey, a story, a shop, and an opportunity through gamification to generate coupons. Here are the noon locations of iinformatica, an innovative SMB and publishing house, become virtualized as a revolutionary e-commerce environment to enhance made in Italy. Each interaction becomes an opportunity for data analysis that takes into account user behavior analysis and empathy/retemption of each interactive content/action/item. All realized following the guidelines of sustainable software development.

References

(1) Osman et al. (2009). Development and Evaluation of an Interactive 360 Virtual Tour for Tourist Destinations, Journal of Information Technology Impact
(2) L. Argyriou et al. (2019), Design methodology for 360° immersive video applications: the case study of a cultural heritage virtual tour, Personal and Ubiquitous Computing

DEFINING A COMPOSITE MEASURE OF EDUCATIONAL POVERTY AT THE INDIVIDUAL LEVEL: A MULTIDIMENSIONAL APPROACH

Antonio De Falco[1], Cristina Davino[2], Rosa Fabbricatore[3], Jonathan Pratschke4, Rosaria Romano[5]

[1] *University of Naples Federico II* (email:antonio.defalco3@unina.it)
[2] *University of Naples Federico II* (email:cristina.davino@unina.it)
[3] *University of Naples Federico II* (email:rosa.fabbricatore@unina.it)
[4] *University of Naples Federico II* (email: jonathan.pratschke@unina.it)
[5] *University of Naples Federico II* (email:rosaria.romano@unina.it)

The concept of Educational Poverty (EP) emerged in the late 1990s within the social sciences and Sen's Capability Approach framework, as the discourse on poverty expanded to encompass various deprivation spheres such as health, housing, and education (1). Since then, EP has been recognised as a significantly impactful form of poverty, particularly among young people. However, despite ongoing efforts, scholars have not yet achieved consensus on its definition and measurement, reflecting the complex nature of this phenomenon.

Starting from these considerations, our work aims to contribute to the debate on the topic by introducing a novel measure of EP at the individual level. By drawing on Save the Children's framework(2), we conceptualised EP as a lack of educational opportunities among youth that undermine their possibility to learn, develop skills, and cultivate aspirations and talents. To operationalise this definition, we employed a set of indicators to assess the educational opportunities available to students across three domains: family, school, and environment. This approach enables us to consider the role played by the different contexts on learning opportunities. Statistically, to define an EP measure, we developed a composite indicator using a Multiple Factor Analysis (3).

In this respect, educational poverty has been conceived as a latent construct measured by three different dimensions (family, school, and territory) and their related set of items. Data collection involved designing a questionnaire to gather comprehensive information on students' family backgrounds, learning opportunities, and educational outcomes such as school performance and socioemotional development. The survey was conducted among a convenience sample of high school students in Naples.

The results contribute to a deeper comprehension of EP, elucidating the influence of diverse contexts on students' learning opportunities. This sheds light on the intricate nature of educational deprivation and its implications for students' cognitive and noncognitive abilities.

References

(1) Sen A, Anand S. Concepts of Human Development and Poverty: A Multidimensional Perspective. In: Poverty and Human Development: Human Development Papers 1997. United Nations Development Programme; 1997. p. 1–20.
(2) Save the Children. La lampada di Aladino. L'indice di Save the Children per misurare le povertà educative e illuminare il futuro dei bambini in Italia. Save the Children; 2014.
(3) Escofier B, Pagès J. Multiple factor analysis. Comput. Stat. Data Anal. 1994;18(1):121–140.

POVERTY AND HIGH SCHOOL STUDENTS' CAREER ASPIRATIONS: A STRUCTURAL EQUATION MODELING APPROACH TO MEASURE AND EXPLORE THE ROLE OF THE CAPACITY TO ASPIRE

Rosa Fabbricatore[1], Antonio De Falco[2], Marco Gherghi[3], Enrica Morlicchio[4]

[1] *University of Naples Federico II* (email: rosa.fabbricatore@unina.it)
[2] *University of Naples Federico II* (email:antonio.defalco3@unina.it)
[3] *University of Naples Federico II* (email:marco.gherghi@unina.it)
[4] *University of Naples Federico II* (email:enrica.morlicchio@unina.it)

The high school period represents the time when young people shape their career aspirations, choosing among several educational or working pathways. Social and material resources significantly affect this process: students from more disadvantaged groups are less likely to attend university and have good labor market outcomes (1). The link between socio-economic backgrounds and career aspirations also encompasses the development of a cultural ability, called the "capacity to aspire" (2), which is defined as the ability to represent the future, to navigate the dense combination of nodes and pathways, to set goals and to make plans to reach them. Consequently, poverty traps may manifest as constraints internal to individuals, such as impaired agency or lack of hope, resulting in a compromised capacity to aspire. Although the role of the capacity to aspire has been recognized as pivotal in enabling aspirations, a clear operational definition of the concept remains elusive in the existing literature, posing challenges for quantitative measurement.
The present work aims to contribute to this research line by (i) proposing an operationalization of students' capacity to aspire along with a set of indicators for its measurement, and (ii) investigating the relationship between students' socio-economic backgrounds and career aspirations, mediated by the capacity to aspire. From a modeling point of view, the capacity to aspire has been conceived as a multidimensional unobserved (latent) variable measured by a set of Likert-type indicators. A structural equation modeling approach (3) has been exploited to validate the factorial structure and assess the strength of the hypothesized relationships. The study involves a convenience sample of high school students living in the Campania region, Italy. The results will reveal the main dimensions of the capacity to aspire, identify those predominantly influenced by socio-economic resources, and underscore its impact on students' career aspirations.

References

(1) Mazenod A, Hodgen J, Francis B, Taylor B, Tereshchenko A. Students' university aspirations and attainment grouping in secondary schools. High. Educ. 2019;78:511–527.
(2) Appadurai A. The capacity to aspire: culture and terms of recognition. In: Vijayendra R, Walton M, editors. Culture and public action. Stanford: Stanford University Press; 2004. p. 59–84.
(3) Bollen KA. Structural equations with latent variables. John Wiley & Sons; 1989.

DETERMINANTS OF VULNERABILITY TO POVERTY: EVIDENCE FROM ITALY

Antonio Acconcia[1], Raffaele Mattera[2], Michelangelo Misuraca[3], Germana Scepi[4], Maria Spano[5]

[1] *University of Naples Federico II* (email: antonio.acconcia@ugr.es)
[2] *University of Rome La Sapienza* (email: Raffaele.mattera@unroma1.it)
[3] *University of Calabria* (email: michelangelo.misuraca@unical.it)
[4] *University of Naples Federico II* (email: germana.scepi@unina.it)
[5] *University of Naples Federico II* (email: maria.spano@unina.it)

Understanding vulnerability to poverty is crucial for shaping effective policy measures and implementing targeted interventions. The literature on poverty, particularly in relation to health and family well-being (2), emphasizes the exploitation of socioeconomic status as a critical factor,
as these are robust predictors of the availability of resources necessary for a decent life. The absence of these resources amplifies vulnerability, leading individuals to experience poverty. In this study, we consider various relevant variables, including the presence of dependent children, the age and education level of the householder, and the household's geographical localization (1). Moreover, we recognize that the analysis of differences in these determinants is relevant, as individuals and communities confront diverse challenges that contribute to their vulnerability to poverty (3). To address this complexity, advanced clustering techniques are employed to identify distinct groups with shared attributes, offering insight into the intricate interplay of factors influencing vulnerability. Notably, the analysis extends its focus to both pre- and post-COVID-19 periods. The global pandemic has introduced new challenges, further emphasizing the need to understand how diverse socioeconomic factors contribute to or mitigate poverty risks in evolving circumstances. By exploring the heterogeneity in determinants across different time frames, this research aims to provide a comprehensive understanding of the dynamics of vulnerability, informing evidence based policymaking and targeted interventions tailored to specific clusters within the population.

References

(1) Acconcia, A., Carannante, M., Misuraca, M., & Scepi, G. (2020). Measuring vulnerability to poverty with Latent Transition Analysis. Social Indicators Research, 151, 1-31.
(2) Blakely, T., Hales, S., Prüss-Üstün, A., Campbell-Lendrum, D. H., Corvalán, C. F., Woodward, A., & World Health Organization. (2004). Poverty: assessing the distribution of health risks by socioeconomic position at national and local levels. World Health Organization
(3) Salvati, L., Zitti, M., & Carlucci, M. (2017). In-between regional disparities and spatial heterogeneity: a multivariate analysis of territorial divides in Italy. Journal of Environmental Planning and Management, 60(6), 997-1015.

ON THE DETERMINANTS OF THE INABILITY OF ITALIAN HOUSEHOLD TO MAKE ENDS MEET

Francesca Condino[1], Filippo Domma[2]

[1] *University of Calabria* (email:francesca.condino@unical.it)
[2] *University of Calabria* (email:filippo.domma@unical.it)

One of the economic-social problems that has emerged in the Italian political debate for several years consists, for an increasingly significant share of families, in the difficulty of making ends meet. A problem that worsened after the COVID-19 pandemic, to the point that EUROSTAT and the National Statistical Institutes have designed a new data collection to investigate the variations in living conditions of the member countries of the European Union, highlighting that ”the share of the population able to make ends meet with great difficulty or with difficulty ranged from 9.2% in Finland to 37.0% in Bulgaria” (1). Here, we study a particular aspect of the so-called financial fragility of Italian households, i.e. the case in which the household disposable income does not cover expected (planned) expenses. As a measure of household fragility we use the probability that non-durable consumption (expected expenditure) is greater than household disposable income.
We specify the joint distribution using copula function to account for the dependence between income and consumption, and a recently proposed reformulation of Dagum distribution to model the marginals. This particular reformulation allows us to express the marginal distributions in terms of indicators of interest. Furthermore, we also reparameterize the copula function, as done for the marginals. To have a tool for evaluating how the heterogeneity among households impact on the measures of interest and on dependence, we relate the marginals and copula parameters to the individual features by choosing suitable link functions, in analogy to generalized linear models. It follows that also the proposed fragility measure depends on the socio-economic characteristics through the specified regressive models. Finally, using data from Survey on Households Income and Wealth (SHIW) by Bank of Italy, we show how the fragility measure varies depending on the socio-economic characteristics of families.

References

(1) https://ec.europa.eu/eurostat/en/web/products-eurostat-news/-/ddn-20220629-1

UNLEASHING THE CREATIVE WAVE: A BENCHMARKING ODYSSEY BETWEEN AI AND GENERATIVE AI IN LANGUAGE TECHNOLOGIES

Federico Neri[1], Maria Caridi[2], Tommaso Petrolito[3]

[1] *Competence Centre of AI applied to Human Language Technologies Deloitte Consulting*
(email: feneri@deloitte.it)
[2] *Competence Centre of AI applied to Human Language Technologies Deloitte Consulting*
(email: mcaridi@deloitte.it)
[3] *Competence Centre of AI applied to Human Language Technologies Deloitte Consulting*
(email: tpetrolito@deloitte.it)

The field of Artificial Intelligence (AI) is undergoing a profound metamorphosis, with high-tech giants revolutionizing the landscape. Amidst this transformation, emerging players are carving their niche, strategically guiding companies to harness the power of data. Yet, the true paradigm shift unfolds with Generative AI (Gen AI), where major tech players like Microsoft/OpenAI, Google, AWS, and Meta lead the charge. Gen AI, mimicking the human creative process, transcends the role of mere enabler, potentially reshaping business models, processes, and societal interactions.

As the global revenues for AI and Gen AI are projected to surpass $900 billion in 2026, with a remarkable 18.6% five-year Compound Annual Growth Rate (CAGR), and a potential $7 trillion boost to Global GDP over a decade, the session offers a comprehensive exploration of the transformative potential of AI and Gen AI. It is not just about technological prowess; it is about reshaping the very essence of how we work, think, and innovate.

This paper is about a captivating journey through the benchmarking odyssey at the forefront of the AI revolution. It delves into the benchmarking of Language Models (LM) and Large Language Models (LLM) applied to human language technologies. Exploring the opportunities and limitations presented by chatGPT, GPT4, Titan/Bedrock, LLAMA-2, Claude-2 and others, the paper showcases groundbreaking AI solutions. Notable examples include Automated Journalism and KGRAIL, Deloitte's AI asset for comprehending and generating language, endorsed by the Deloitte AI Institute and acknowledged globally. The session unveils innovations in algorithmic design, addressing challenges like hallucination, multi-linguism, and cross-linguism, demonstrating Deloitte's commitment to pushing the boundaries of language-based AI technologies

CO-OCCURRENCE NETWORK TO EXPLORE RESEARCH TOPICS EVOLUTION AMONG ITALIAN STATISTICIANS

Amin Gino Fabbrucci Barbagli[1], Domenico De Stefano[2], Francesco Santelli[3], Susanna Zaccarin[4]

[1] *University of Trieste* (email: amingino.fabbruccibarbagli@phd.units.it.)
[2] *University of Trieste* (email: ddestefano@units.it)
[3] *University of Trieste* (email: fsantelli@units.it)
[4] *University of Trieste* (email:susannaz@deams.units.it)

Research interests in scientific disciplines evolve over time due to new ideas, new theoretical and applied issues, as well as improved methods and analytical tools. The dynamics of specific disciplines can be retrieved by looking at the terms and words occurring in the scholars' published works in order to point out evergreen topics or emerging ones. We focus on the 836 Italian academic Statisticians - grouped into five research subfields consisting in Statistics, Statistics for Experimental Research, Statistics for economic and social applications, and Demography - as they are listed in the official *MUR Cineca* database at the end of 2022. We retrieved their published works on *Scopus* in the last ten years since 2012. The database, made up of 12485 papers (including references, abstracts, keywords, and the number of citations per work), has been analyzed to answer the following research questions: *a)* are the different Italian subfields consistent in their production, that is do the topics they covered clearly differ among them? *b*) are there trending topics overall? Are they changing over time? *c*) are the techniques used by statisticians in their works evolving/modifying over time? *d*) what are the characteristics of the most successful contributions? Are they linked to some specific topics or scholars? To answer these questions we perform co-occurrence analyses of keywords and abstracts, in order to find communities describing scientific subjects of interest, and topic extractions from the textual data. Furthermore, analyses will be presented over time to explore the longitudinal dynamics of the scientific production and topics by applying Latent Dirichlet Allocation (LDA) and Social Network Analysis tools.

AI REVOLUTION IN HEALTHCARE AND MEDICINE: TRANSFORMATIVE INSIGHTS TROUGH NATURAL LANGUAGE PROCESSING

Alessia Forciniti[1], Francesco Santelli[2]

[1] I*ULM University* (email: alessia.forciniti@iulm.it)
[2] *University of Trieste* (email: fsantelli@units.it)

The complexity and increase of data in the health sector imply that artificial intelligence (AI) and related technologies are becoming an integral part of life sciences in administrative aspects, in patients' engagement, and for diagnosing and treating diseases. The related scientific literature to date discusses the advantages, methodological applications, future challenges, and ethical issues.
To understand the extent to which AI has developed in healthcare and detect insights into the transformative pillars, we conducted a content analysis on academic literature from 2000 to 2023. The corpus comprises 10,775 documents obtained from "Web of Science". We conducted a data integration of two extractions of abstracts containing the terms "artificial", "intelligence", and "healthcare" in the first extraction and "clinical" in the second one. A natural language processing approach was used, focusing on two research directions: 1) detecting the most relevant topics and 2) classifying the documents regarding the main domain application categories of AI identified in the first step. For the first aim, we performed on corpus a machine learning model based on topic modeling in embedding spaces called ETM that allowed us to determine two distinct semantic pillars: diagnosis and treatment. Based on the identification of the topics, a semi-automatic annotation of the two semantic pillars was achieved for the whole corpus.
To deepen the semantic dimension of ETM and our categories annotation, the second step focused on a comparative classification strategy. Two approaches were adopted, one based on the decision tree rules (random forest (RF)) and the other on the neural network approach (long short-term memory (LSTM)).
The findings suggest that a clear difference emerges between the two domains, with just a marginal quota of scientific papers encompassing both but with few notable emerging topics, such as e-health, AI-assisted meta-analyses, and clinical decision support systems.

STATISTICAL ANALYSIS OF VISITORS' ONLINE REVIEWS FOR ARTISTIC AND CULTURAL ATTRACTIONS

Riccardo Ricciardi[1], Marica Manisera[2], Paola Zuccolotto[3]

[1] *University of Brescia* (email: riccardo.ricciardi@unibs.it)
[2] *University of Brescia* (email: marica.manisera@unibs.it)
[3] *University of Brescia* (email: paola.zuccolotto@unibs.it)

This contribuGon falls within the research field of the statistical analysis of grouped text documents and focuses on developing a language model trained on online reviews posted on Google by visitors of the main cultural a2racGons in Brescia, Italy. These reviews encompass four primary cultural desGnaGons managed by Fondazione Brescia Musei: the Castle, the picture gallery Pinacoteca Tosio-MarGnengo, the Roman Brixia archaeological site, and the Santa Giulia Museum. The primary objecGve is to create a language model capable of semanGc representaGon and categorizaGon of reviews into the four disGnct areas defined by the mentioneoned cultural a2racGons. The objecGve is achieved through the fine-tuning of a BERT model [1] for a multiclassifiction task. The proposed approach involves iniGally designing a classificaGon model using reviews where the a2racGon is known. The model's key uGlity, however, lies in its ability to idenGfy a2racGons in text documents, covering both reviews and non-reviews. This is particularly valuable in cases where documents lack explicit labels identifying the cultural a2racGon they reference.
The model streamlines the task of comprehending and categorizing reviews, eliminaGng the need for laborious manual reading. By processing and classifying texts, the proposed model facilitates the expansion of the online discourse database related to cultural a2racGons, not limited to Brescia.
The proposed model offers an efficient automaGc soluGon to the challenge of comprehending and categorizing reviews associated with cultural a2racGons. This approach not only saves Gme and effort but also contributes to a more efficient organizaGon of vast textual data and provides valuable insights into public opinions about these a2racGons across various online plaZorms.

References

(1) J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirecGonal transformers for language understanding,” arXiv preprint arXiv: 1810. 04805, 2018.

Acknowledgements

The methodological part of this study was carried out within the MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from Next- GenerationEU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1551.11-10-2022, PE00000004).
The applicaNon in this study has been supported by Fondazione Cariplo, grant n° 2020-4334, project Data Science for Brescia (bodai.unibs.it/ds4bs/), Big & Open Data InnovaNon Laboratory (BODaI- Lab, h9ps://bodai.unibs.it/).

UNVEILING THE SOCIO-ECONOMIC IMPACTS: A ROBOTIC PROSTHETIC HAND PROJECT

Sara Preti

University of Genoa (email: sara.preti@edu.unige.it)

The concept of social impact has garnered increasing attention in recent years, particularly through social impact assessments aimed at identifying and emphasizing the added value and social changes brought by project activities, alongside the sustainability of social actions. This study aims to examine the economic and social impact of a robotics research project culminating in the development of Hannes, an advanced prosthetic hand resulting from collaboration between IIT and INAIL. Hannes is utilized in robotic assistance for individuals who have undergone upper limb amputation due to physical trauma or surgery.

Data were gathered through a clinical study involving patients using myoelectric prostheses, who underwent functional and psychosocial assessments before and after utilizing Hannes. The Social Return On Investment (SROI) methodology will be employed in this study to evaluate the impacts of the robotic device. This methodology delves into the intricate relationships between qualitative, quantitative, and financial information, analyzing them to estimate the value created by research activities.

Preliminary results indicate that for every euro invested in the project, approximately 9 euros of social value are generated. This revelation highlights the substantial positive impact of Hannes and underscores the potential for further advancements in the field of robotics to enhance the lives of individuals with limb loss. Through comprehensive analysis, this research contributes to a deeper understanding of the societal benefits arising from innovative technological solutions in healthcare that mitigate the social inequalities.

SKELLAM REGRESSION FOR DEATH COUNT MODELLING

Giacomo Lanfiuti Baldi[1], Andrea Nigri[2]

[1] *La Sapienza University of Roma* (email: giacomo.lanfiutibaldi@uniroma1.it)
[2] *University of Foggia* (email: andrea.nigri@unifg.it)

Modelling mortality stands as a fundamental tool in understanding demographic patterns and projecting future trends. Mortality data are often reported as count data within specific age-period domains for various populations. The simplest models for mortality patterns typically involve age and period as the sole explicative variables, since we do not need any more information than what is intrinsically contained in the data. This approach forms the cornerstone for exploring and comprehending the intricate dynamics of mortality across diverse populations. Furthermore, our focus lies in delineating the disparities between two populations—these populations could manifest as two genders, distinct countries, or differing causes of mortality. We aim to provide a statistical framework for jointly modelling the mortality of two populations, beginning with the shared assumption that the count of deaths at each age and year follows a Poisson distribution.
This study proposes an innovative approach to comparing deaths due to two causes, integrating the Skellam distribution in an Age-Period model. We contrast this methodology with conventional models employing two independent Poisson distributions or a Bivariate Poisson distribution to model the two causes. We present results based on both simulated and real data.

SPATIAL CLUSTERING ALGORITHMS FOR THE MULTIDIMENSIONAL ANALYSIS OF SOCIAL INEQUALITIES

Corrado Crocetta[1], Leonardo Salvatore Alaimo[2], Paola Perchinunno[3], Samuela L'Abbate[4]

[1] *University of Bari "Aldo Moro" (*email: corrado.crocetta@uniba.it)
[2] *University of Rome La Sapienza (*email: leonardo.alaimo@uniroma1.it)
[3] *University of Bari "Aldo Moro" (*email: paola.perchinunno@uniba.it)
[4] *University of Bari "Aldo Moro" (*email: samuela.labbate@uniba.it)

The analysis of social inequalities is a topic of current interest and is studied as a factor in the evolution and measurement of the level of well-being. A fundamental prerequisite for a correct statistical analysis of this phenomenon is the need to share a univocal definition of the concept of social sustainability.
This work starts from the need to identify territorial areas and/or population subgroups characterized by situations of hardship or strong social exclusion through the construction of indicators that can estimate situations of social inequalities in small areas.
Scientific research options have been oriented towards the establishment of a multidimensional approach, sometimes renouncing dichotomous logic to go as far as fuzzy classifications in which each unit simultaneously belongs and does not belong to the selected category. In this work, multidimensional statistical analysis methodologies (TFR method) and territorial clustering methods will therefore be used to aggregate adjacent spatial units with high intensity of the phenomenon (DBSCAN and Seg-DBSCAN).

LEVERAGING THE MULTIWAY APPROACH FOR COHORT GENDER GAP ANALYSIS

Susanna Levantesi[1], Pietro Giordani[2], Andrea Nigri[3]

[1] *University of Rome La Sapienza (*email: susanna.levantesi@uniroma1.it)
[2] *University of Rome La Sapienza (*email: paolo.giordani@uniroma1.it)
[3] *University of Foggia(*email: andrea.nigri@unifg.it)

Understanding mortality is of great importance for both private and public sectors to design appropriate pension or insurance plans. To this purpose, several interesting applications of multi-way models to mortality data are available in the literature (1). Generally speaking, in these studies, data usually refer to mortality rates across demographic features such as causes of death, ages, countries, and years. This work represents a further step in mortality analysis by focusing on the gender gap (2) in causes of death and its evolution by cohort. Limiting our attention to the three-way case, the Tucker3 model is applied to a tensor containing gender gap data in mortality distinguished by causes of death, age classes, and cohorts.

References

(1) CARDILLO, G., GIORDANI, P., LEVANTESI, S., NIGRI, A., & SPELTA, A. 2023. Mortality forecasting using the four-way CANDECOMP/PARAFAC decomposition. Scandinavian Actuarial Journal, in press.
(2) ZARULLI, V., KASHNITSKY, I., & VAUPEL, J.W. 2021. Death rates at specific life stages mold the sex gap in life expectancy. Proc. Natl. Acad. Sci.

EXPLORING INTERNATIONAL DATA ON STUDENTS' READING PERFORMANCE WITH S-CONCORDANCE MEASURES

Simona Korenjak-Černe

University of Ljubljana (email: Simona.Cerne@ef.uni-lj.si)

Measures s-concordance and s-discordance were introduced in 2019 by E. Diday (1) to measure the agreement (or disagreement) between an object and a collection of objects. The prefix "s" refers to symbolic data analysis, in which the descriptions of the objects, that represent aggregations of individuals and are referred to as classes, preserve more internal variability of the individuals, which is also further taken into account in the analysis.
A high s-concordance of a class “c” with a given collection of classes “P” for a category “x” reflects the high frequency of this category among the individuals within the class “c” (high “internal frequency”) in combination with the additional information, that there are numerous classes in the given collection of classes that have approximately the same internal frequency for this category “x”. On the other hand, a high s-discordance characterizes a very frequent category “x” in an observed class “c” (high “internal frequency”) that is specific to it, which means that there are not many classes with such an internal frequency. More theoretical background can be found in the recently published chapter (2).
In our study, we applied selected s-concordance and s-discordance measures to the PIRLS dataset (3). The PIRLS dataset is based on a large-scale international evaluation of student performance in traditional paper reading and reading on digital devices in several countries around the world. Measuring s-concordance and s-discordance gives us additional insight into a country's reading achievements compared to the collection of countries. Similarly, it can provide additional information about the reading achievements of individual classes within a country. Furthermore, we explore how these new methodological approaches could enable a more comprehensive comparison between traditional reading on paper and reading on digital devices, the importance of which has increased enormously in the "Covid era".

References

(1) Diday E. Concordance and discordance between classes of complex data. Presented at the Workshop Advances in Data Science for Big and Complex Data: From data to classes and classes as new statistical units, University Paris-Dauphine, January 10-11, 2019.
(2) Diday E. Introduction to the “s-concordance” and “s-discordance” of a Class with a Collection of Classes. In: Beh EJ, Lombardo R, Clavel JG, editors. Analysis of Categorical Data from Historical Perspectives. Behaviormetrics: Quantitative Approaches to Human Behavior, vol 17. Springer, Singapore, 2023. https://doi.org/10.1007/978-981-99-5329-5_27
(3) PIRLS Progress in International Reading Literacy Study. https://timssandpirls.bc.edu/pirls2016

RECOGNITION OF EMOTIONS BY S-DISCORDANCE MEASURE: SUPERVISED AND UNSUPERVISED APPROACH

Jasminka Dobša

University of Zagreb (email: jasminka.dobsa@foi.unizg.hr)

The aim of this research is to apply s-discordance measure introduced by E. Diday (1) in the context of classification of textual documents according to emotion. For that purpose, weight of word or index term will be measured by s-discordance measure S_{disc} (c, P, x) of a class of documents c with collection of all classes P for a given index term x. Classes of documents are created based on the emotion present in the document (anger, disgust, fear, guilt, joy, sadness, and shame). S-discordance measure defined for this application is measuring relevance of the index term x for a class of documents or, for mentioned emotions. The relevance of term x for a class c is high if the proportion of documents inside the class c that contain that term is high, and the number of classes for which proportion of documents in that class that contain term x is higher than in class c is low. Measuring of word emotions by s-discordance measure will be applied in two different settings: in a supervised way by extension of standard Tf-Idf weighting scheme by s-discordance measure, and in an unsupervised way by automatic creation of lexicon of emotions using s-discordance measure of index terms as their weights and classification of test documents using lexicon. For experiment will be used ISEAR data set (2).

References

(1) Diday, E. (2020) Explanatory tools for machine learning in the symbolic data analysis framework. In Diday, E. Guan, R., Wang, H. (eds.) *Advances in Data Science,* ISTE-Wiley.
(2) International Survey on Emotion Antecedents and Reactions “, ISEAR, data set and description of the questionnaire, data treatment, and variable abbreviations as used in the data base, https://www.unige.ch/cisa/research/materials-and-online-research/research-material/, accessed January 2024

VISUALIZING MULTIDIMENSIONAL DISTRIBUTIONAL DATA: SOME NEW TOOLS

Antonio Irpino

University of Campania "L. Vanvitelli" (email: antonio.irpino@unicampania.it)

Visualization tools are crucial for conveying patterns and guiding analytical approaches in data interpretation. When faced with the complexity of visualizing distributional data tables—where each observation is a vector of frequency or density distributions—the need for user-friendly tools becomes apparent. To address this challenge, three innovative visualization tools have been introduced for data tables characterized by numeric distributional data.

The first two tools, the Green Eye Iris (GEI) and the Flower plot, utilize a polar coordinate-based representation of stacked bar charts or violin plots.

The third tool extends the traditional heatmap plot and is particularly effective for illustrating datasets with numerous observations and variables. All three methods focus on visually representing the proportion of mass distributed on the domain variable, utilizing diverging color palettes for each distribution.

References

(1) Brito, P., and S. Dias. 2022. Analysis of Distributional Data. Chapman; Hall/CRC. https://doi.org/https://doi.org/10.1201/9781315370545.

EDUCATIONAL DATA SCIENCE: CHALLENGES AND OPPORTUNITIES IN A RAPIDLY EVOLVING INFORMATION AGE

Clelia Cascella[1], Maria Pampaka[2]

[1] *INVALSI, The University of Manchester* (email: cascella.cecilia@gmail.com)
[2] *University of Manchester* (email: maria.pampaka@manchester.ac.uk)

Data Science (DS) has become widespread in various fields of knowledge. However, in the field of Education, the applications of DS have been limited. Although it has been a decade since these applications have become more frequent, it has been argued that there is a lack of a reference community, society and journal to guide and help disseminate the work done on Educational Data Science (EDS) (1). Our proposed paper aims to contribute to this debate by presenting the state-of-the-art in EDS through a systematic literature review. We will contribute to the discussion by (i) providing an overview of existing definitions of EDS, and (ii) exploring/discussing the features of EDS (especially with regard to data, methods and ethics).
We will use the PRISMA model (2) to systematically review the literature. Our selected keywords will be used in the main databases to search for both theoretical and empirical studies. Both scientific publications (i.e. peer-reviewed journal articles and books) and institutional reports will be included, without any language or time restrictions, to find all relevant publications. From these records, a snowballing search will be carried out to complete the literature search. Based on the results of our analysis, we will try to understand if EDS should be considered as a new, emerging discipline or just as an umbrella term. Furthermore, by including both scientific and grey literature, we aim to represent and bring together different perspectives on EDS, namely those of researchers, educators and practitioners. Working at the intersection of DS and education is rapidly becoming the next frontier of educational research (3). Understanding the features of EDS (and how it differs from other disciplines), its challenges and affordances, must therefore be seen as a priority to beter inform both policy and practice.

References

(1) Peña-Ayala, A. (2023). “Educational Data Science: An “Umbrella Term” or an Emergent Domain? In Educational Data Science: Essentials, Approaches, and Tendencies: Proactive Education based on Empirical Big Data Evidence”. Singapore: Springer Nature Singapore.
(2) Welch, V., Petticrew, M., Petkovic, J., Moher, D., Waters, E., White, H., ... & Wells, G. (2016) Extending the PRISMA statement to equity-focused systematic reviews (PRISMA-E 2012): explanation and elaboration. Journal of Clinical Epidemiology, 70, 68-89.
(3) McFarland, D. A., Khanna, S., Domingue, B. W., & Pardos, Z. A. (2021). Education data science: Past, present, future. AERA Open, 7, 23328584211052055.

DATA ETHICS IN THE OPEN: NAVIGATING STUDENT INFORMATION RISKS IN EDUCATIONAL SOCIAL MEDIA

Joshua M. Rosenberg[1], Conrad Borchers[2], Macy Burchfield[3], Christian Fischer[4], Sondra Stegenga[5]

[1] *University of Tennessee* (email: jrosenb8@utk.edu)
[2] *Carnegie Mellon University* (email: cborcher@andrew.cmu.edu)
[3] *University of Tennessee*
[4] *University of Bolzan* (email: christian.fischer@uni-tuebingen.de)
[5] *University of Utah* (email: sondra.stegenga@utah.edu)

In an era of ubiquitous social media use, educational institutions leverage social media platforms to engage with their communities and stakeholders. While there are many benefits related to how educational institutions can engage with social media, this trend also raises serious ethical considerations, primarily regarding the large-scale dissemination of students' personally identifiable information (PII). Using large data sets exacted from Facebook and X (formerly and colloquially known as Twitter), we report on investigations of the social media practices of schools and school districts in the United States and the extent to which they potentially put students' PII at risk of datafication on public social media pages (1). Furthermore, we examine the ethical ramifications related to the data mining of social media data for educational purposes. Mainly, how should we as researchers, approach using the social media data that we personally do not think should be available to the public? How does the use of this data intersect with locally-governed institutional review boards (IRB) that often consider it exempt from human-subject research because it is publicly available data? And what are the roles and responsibilities of educational practitioners and parents? We relate our discussion to debates about data ethics and the datafication of students' information.

References

(1) Rosenberg, J. M., Borchers, C., Burchfield, M. A., Anderson, D., Stegenga, S. M., & Fischer, C. (2022). Posts about students on Facebook: a data ethics perspective. Educational Researcher, 51(8), 547-550.

WHAT TIMESCAPE FOR EDUCATIONAL DATA? SLOWNESS, TEMPORAL CARE AND AN ETHICS OF THE POSSIBLE

Emiliano Grimaldi[1], Jessica Parola[2]

[1] *University of Naples Federico II (*email:emiliano.grimaldi@unina.it)
[2] *University of Naples Federico II* (email: jessica.parola@unina.it)

This presentation contributes to a collective reflection on what kind of educational data and what kind of educational data science we do need if we want to enable teachers and educators in their search for a fair, inclusive, democratic and non-anticipatory education. To do so, we move from a critical discussion of the key features of the contemporary data imaginary, which presents data as *fast*, *accessible*, *panoramic*, *revealing*, *prophetic* and *smart* (1). First, we show how this imaginary, and the related infrastructure, are playing a major role in shaping the understanding and uses of data and data analytics in the educational domain (6) and are contributing to a re-framing of the temporal rhythms, relations and modalities in/of education. Second, we problematize such an emerging timescape (4) for both educational data and the associated data uses, shedding light on the risks that they entail for education, which we relate to the reinforcement of a determinist, reductionist and discriminatory pedagogy (5, 3). Drawing on Kitchin (3) and Biesta (2), we reflect on the potential for a different timescape for educational data that is articulated around *slowness, temporal care* and an *ethics of the possible*. We suggest how this requires to think, produce and use educational data in a perspective that: a) widens the spaces for deceleration, disconnection and asynchronicity; b) enables the cultivation and valuing of multiple and overlapping educational temporalities, histories and unfoldings; and c) includes the possibility of forgetting and a democratization of future-making.

References

(1) Beer, D. (2019). *The Data Gaze: Capitalism, Power and Perception*. London: Sage.
(2) Biesta, G. J. (2016). *The Beautiful risk of education*. London: Routledge.
(3) Grimaldi, E. (2019). *An archaeology of educational evaluation: Epistemological spaces and political paradoxes*. London: Routledge.
(4) Kitchin, R. (2023). *Digital timescapes: Technology, temporality and society*. Cambridge: Polity Press.
(5) Jandrić, P., & Hayes, S. (2022). Postdigital critical pedagogy. In *The Palgrave handbook on critical theories of education* (pp. 321-336). Cham: Springer International Publishing.
(6) Williamson, B. (2017). *Big Data in Education: The Digital Future of Learning, Policy and Practice*. London: Sage.

SYNTHETIC POPULATIONS AND AGENT-BASED MODELING: CHALLENGES AND PROSPECTS IN THE OPEN SCIENCE

Rocco Paolillo[1], Mario Paolucci[2]

[1] *National Research Council of Italy, Institute for Research on Population and Social Policies (CNR IRPPS)* (email: rocco.paolillo@cnr.it)
[2] *National Research Council of Italy, Institute for Research on Population and Social Policies (CNR IRPPS*) (email: mario.paolucci@cnr.it)

Agent-based modeling allows studying the emergence of collective phenomena such as segregation or polarization through the interaction of virtual agents simulating human behaviors and cognitions (1) To ensure the reliability of results from agent-based models, synthetic populations aim at building agents that are representative of the target population. The process can include the synthetic reconstruction of socio-demographic characteristics of agents manipulating data for initialization or integrating independent sources of data. Research in the field is devoted to producing algorithms for the extraction of synthetic populations, most of them applying statistical techniques to data archived, such as the Iterative Proportional Fitting and its extensions to spatial scales and nested data.[2] The availability of data infrastructures for social data science nowadays allows for the integration of different data sources, increasing the potential of synthetic populations for policy testing and social research within the principles of Open Science. However, this comes with some challenges peculiar to agent-based modeling. First, developing self-made algorithms to extract synthetic populations can be cumbersome work requiring efforts that distract from the scope of the simulation and that increase the risk of committing errors. Additionally, synthetic reconstruction requires familiarization with the data collection and data management of independent sources to harmonize data and scales. To facilitate agent-based modelers against these challenges, the Fostering Open Science in Social Science Research project (FOSSR) envisages an automated service to extract synthetic populations from an integrated database infrastructure. Our contribution first shows the state-of-the-art of literature and methods for synthetic populations, highlighting the challenges specific to agent-based modeling, and how they are dealt with in the development of a software as a service within the FOSSR open-cloud.

References

(1) Macy, M., Willer, R. From factors to actors: computational sociology and agent-based modeling. Annual Review of Sociology. 2002 Aug;28(1), 143-166.

OPEN SCIENCE FOR SOCIAL IMPACT: THE FOSSR POLICY LEARNING PLATFORM

Giovanni Cerulli[1], Andrea Orazio Spinello[2]

[1] *CNR-IRCrES - Research Institute for Sustainable Economic Growth* (email: giovanni.cerulli@ircres.cnr.it)
[2] *CNR-IRCrES - Research Institute for Sustainable Economic Growth* (email: andrea.spinello@ircres.cnr.it)

The advancement of Open Science is greatly encouraged by the establishment and strengthening of Research Infrastructures (RIs). Within the realm of Social Sciences, the promotion of Open Science through RIs aims to foster a dynamic and productive diffusion of knowledge among scholars and between science and society. Beyond integrating and sharing scientific data, RIs can provide innovative research tools and services, foster communitybuilding initiatives and engage with social and political stakeholders to promote an increasingly meaningful impact of scientific work in the choices that affect citizens.
The FOSSR project, funded by NRRP and managed by CNR, is committed to creating a social and research environment that enables simplified and shared access to social science data from three European RIs (CESSDA, RISIS and SHARE) through innovative interfaces and services. A central aspect of the FOSSR project is the creation of advanced tools and services for data collection and analysis to enhance the effectiveness of data-driven socioeconomic policy learning. Of particular significance is the development of a Policy Learning Platform (PLP), based on the latest developments in machine learning and artificial intelligence. The PLP will play an important role in bridging the gap between recent theoretical developments in socioeconomic policy learning and their actual implementation in the actual policy context. The platform is designed with an open source approach and will be developed using three software tools: two open source options, Python and R, and one commercial tool, Stata. The PLP aims to provide researchers, practitioners, and policy makers with the ability to predict policy effects and devise targeted scenarios across a wide range of social sectors. It aims to become a relevant resource in promoting the connection between theory and practice in data-driven policy learning within the context of open science, serving both scientific and social impact objectives.

References

(1) Farago, P. (2014). “Understanding How Research Infrastructures Shape the Social Sciences: Impact, challenges, and outlook”, in Duşa, A., Nelle, D., Stock, G. and Wagner, G.G., Facing the Future: European Research Infrastructures for the Humanities and Social Sciences, SCIVERO Verlag, Berlin, pp. 21-33.
(2) Hallonsten, O. (2020). Research infrastructures in Europe: The hype and the field. European Review, 28(4), 617-635.
(3) Cerulli, G. (2023). Optimal treatment assignment of a threshold-based policy: empirical protocol and related issues, Applied Economics Letters, 30, 8, 1010-1017.

JURASSIC GUIDE: INNOVATIONS AND CHALLENGES TO SURVEY CHILD WELL-BEING IN ITALY

Daniela Cocchi[1], Giulio Ecchia[2], Francesco Giovinazzi[3], Chiara Monfardini[4], Francesca Tosi[5], Massimo Ventrucci[6], Matthew John Wakefield[7]

[1] *Alma Mater Studiorum University of Bologna* (email: daniela.cocchi@unibo.it)
[2] *Alma Mater Studiorum University of Bologna*
[3] *Alma Mater Studiorum University of Bologna*
[4] *Alma Mater Studiorum University of Bologna*
[5] *Alma Mater Studiorum University of Bologna*
[6] *Alma Mater Studiorum University of Bologna*
[7] *Alma Mater Studiorum University of Bologna*

In this paper we discuss the methodological innovations and challenges related to surveying child wellbeing in Italy in the framework of the GUIDE (Growing up in Digital Europe) Research Infrastructure. GUIDE is an ESFRI 2021 Roadmap research infrastructure and aims to develop a comparative longitudinal survey on child well-being in 20 European countries thanks to an accelerated cohort design: in particular, an infant cohort (followed from age 1 to 24 years) and a children cohort (followed from 8 to 24 years). The data collected through this longitudinal survey (according to a FAIR approach) will provide information about a holistic evaluation of child well being and will constitute the basis for public policies at different territorial level in each of the participating countries as well as creating a unique dataset for comparative policy analysis. Taking stock from the pilots conducted in five countries in 2023, the GUIDE Italian team based at the University of Bologna has pre-tested the implementation of the survey in Italy and has produced, in collaboration with CNR-IRPPS, a design for the pilot to be conducted in Italy in the spring 2024. This design allows also to evaluate different survey modes and their implication for data quality (that is comparing CAPI and CAWI survey methodologies) and to test the robustness of the results with regard to regional heterogeneity, that is comparing northern regions (Lombardy and Emilia Romagna) with southern regions (Campania and Apulia), among other more specific technical elements of the sampling strategy. This paper analyses the work conducted so far and presents pathways to the full-scale launch of the GUIDE survey in 2026 in Italy

PLAY IS THE NEW STAT

Rina Camporese[1], Susi Osti[2], Monica Bailot[3], Enrico Caleprico[4], Maria Marino[5]

[1] *Istat - Italian National Institute of* Statistics (email: rina.camporese@istat.it)
[2] *Istat - Italian National Institute of* Statistics (email: susi.osti@istat.it)
[3] *Istat - Italian National Institute of* Statistics (email: monica.baillot@istat.it)
[4] *Istat - Italian National Institute of* Statistics (email: enrico.caleprico@istat.it)
[5] *Istat - Italian National Institute of* Statistics (email: maria.marino@istat.it)

The concept "statistical literacy" is often restricted to a well-defined area: statistics at school with enhanced teaching methods. Such a context usually sees statisticians in the asymmetric role of experts while all the others are rewarded when they acquire formal statistical skills. This contribution embraces a wider approach and proposes reflections and experiences for a broader view. In addition to traditional teaching activities, the authors reflect on passing warp and weft statistical threads throughout the entire cultural fabric. Their vision of statistical dissemination covers a broad conceptual spectrum and therefore it is possible to talk of "statistical culture". Culture is not classically referred to as individual training (prescriptive: having to be, having to know), but as used by current social sciences: behaviours that transcend the cultured/uncultured distinction, i.e. systems of ideas and symbols that pervasively guide individual behaviour and thinking (opportunities to experience). Therefore, the learning approach changes from "let's teach it better", to "let it emerge in social life and involve [young] people in it". The challenging journey envisioned, not only foresees a change in the skills of non-statisticians, but also – and perhaps more so – in the attitudes of both statisticians and statistics. In this process, playful activities *play* a fundamental role: they provide a relaxed and appealing setting where a topic like statistics – generally feared or criticized – becomes a friendly experience for everyone and for the young in particular. Before tackling formal statistical language, concepts become game and role-playing experiences, and they are actively lived through collaborative joyful moments, where everyone can contribute, despite their previous knowledge or formal expertise. Two such experiences are presented: one suited for schools from kindergarten to high school, the other open to all citizens.

SCULPTING TOMORROW: EXPLORING FUTURES STUDIES AND STATISTICS THROUGH GAMIFICATION

Simone Di Zio

University "G. d'Annunzio" (email: s.dizio@unich.it)

This review explores the landscape of games within the realm of futures studies, emphasizing their significance in fostering engagement, learning, and strategic foresight. The proliferation of games addressing future scenarios reflects a growing recognition of their potential to democratize futures thinking and involves diverse stakeholders in envisioning possible futures.

The review categorizes existing games according to their objectives, mechanics, and intended audience. From serious games designed for academic and professional training to educational board games targeting a broader audience, the spectrum of futures-focused games is diverse. The analysis encompasses digital simulations, scenario-based board games, and participatory activities that harness the power of play to facilitate understanding and anticipation of future developments.

One key aspect under scrutiny is the pedagogical value of these games. Many offer interactive learning experiences that transcend traditional educational methods, allowing participants to actively engage with complex concepts such as scenario planning, strategic foresight, and decision-making under uncertainty. The integration of gaming elements enhances participant motivation, making futures studies more accessible and enjoyable.

Also in the realm of mathematics and statistics, a plethora of games exists, ranging from board games to digital applications, offering dynamic and enjoyable ways to explore and apply quantitative concepts.

However, in almost all cases, games developed for educational and training purposes are mono-sectoral and do not address the interdisciplinarity that is fundamental to addressing society's problems. Indeed, to date, there is a lack of games that effectively integrate the development of both statistical and mathematical skills, which are predominantly quantitative, with aspects of futures studies, which are primarily qualitative. Understanding and applying probability - which is essential for navigating uncertainty - can play a proactive role in exploring and using both one's personal futures and societal futures, providing the right tools for discriminating between multiple futures, ultimately leading to statistical decision-making.

Recognizing the enormous potential of games for the development of complex skills, it would be extremely helpful and appropriate the creation of such a game, in order to foster an interdisciplinary and holistic approach to skill development, integrating statistical, mathematical, and futures studies capabilities.

BUSINESS STATOOLS: AN INNOVATIVE TOOL FOR STATISTICS LEARNING

Vito Santarcangelo[1], Antonio Ruoto[2], Angelo Romano[3], Saverio Gianluca Crisafulli[4], Luigi Fabbris[5]

[1] *iInformatica srl* (email: vito@iinformatica.it)
[2] *iInformatica srl* (email: antonio@iinformatica.it)
[3] *iInformatica srl*
[4] *iInformatica srl*
[5] *Tolomeo Studi e Ricerche* (email: luigi.fabbris@unipd.it)

This research paper aims to present Business Statools, a card game, a new educational tool for statistics learning and giving appeal to statistics. The game aims to improve the ability of players to pair business targets and statistical techniques. The tool consists of two packs of cards and a die. The players are asked to pick a number of cards up from the statistical techniques pack, according to the tossed die, and keep the best set of technique-cards that satisfy the randomly drawn business object card. The game purpose is to help players to learn statistics through card selection and between-player discussion. More and more young people are fascinated by terms such as artificial intelligence or machine learning, totally ignoring the fact that statistics often has little appeal behind such terms. Through the discovery of business goals and the search for methods to pursue them, statistics becomes a game and thus becomes a time for sharing and growth. The game was created as an evolution of the territorial educational path of the experimental publishing projects Lucanum and Robocom.

THE DEFINITION OF SUSTAINABILITY FOR ITALIAN STAKEHOLDERS: EVIDENCES FROM A SURVEY

Raffaele Angelone[1], Paolo Mariani[2], Andrea Marletta[3], Mariangela Zenga[4]

[1] *University of Milano-Bicocca* (email: raffaele.angelone@unimib.it)
[2] *University of Milano-Bicocca* (email:paolo.mariani@unimb.it)
[3] *University of Milano-Bicocca* (email:andrea.marletta@unimb.it)
[4] *University of Milano-Bicocca* (email: mariangela.zenga@unimib.it)

The concept of sustainability has been spread like one of the most cited and interesting trend topics giving a lot of definitions and good practices to perceive it. In this context, the perception of sustainability for the population and the companies is one of the aspects less frequently faced. In this work, this issue has been analysed using results from a survey in which a sample of 2,000 respondents answered to the perception of sustainability for Italian companies. The main aim of the work is to detect to measure the sustainability concept level of understanding and the perception of the role played by the companies in this field. This objective was achieved asking the population how much they are informed about sustainability, SDGs and how much they are available to pay to choose a sustainable company respect to one not sustainable. Segments of respondents have been achieved using decision trees. The decisional rule used to obtain the nodes was the CHAID (Chi-squared Automatic Interaction Detection) method. Trees are connected acyclic graphsfundamental to create data structures, classification tools, decision theories and prediction models. Each node can be thought of as a cluster, tree prediction models add two ingredients: the predictor and predicted variables labeling the nodes and branches. The proposed approach divided the respondents in four groups showing four different behaviors towards the sustainability. The survey highlights the need for more widespread information to create a widely disseminated culture of sustainability that appears as a concept not yet fully defined.

MIND THE GENDER GAP: EXPLORING INCLUSIVITY IN THE ITALIAN LIFE SCIENCES COMPANIES

Laura Benedan[1], Cinzia Colapinto[2], Paolo Mariani[3], Laura Pagani[4], Mariangela Zenga[5]

[1] *University of Milano-Bicocca* (email: laura.benedan@unimib.it)
[2] *University of Venice* (email: cinzia.colapinto@unive.it)
[3] *University of Milano-Bicocca* (email:paolo.mariani@unimb.it)
[4] *University of Milano-Bicocca* (email: laura.pagani@unimib.it)
[5] *University of Milano-Bicocca* (email: mariangela.zenga@unimib.it)

Empowering all individuals, irrespective of their gender, leads to diverse perspectives, innovation, and inclusive growth. Recognizing the importance of diversity and fostering an environment where everyone has an equal opportunity to thrive, sets the foundation for a future that is both sustainable and equitable.

Over the years, dedicated efforts and specific policies have played a pivotal role in advancing gender equality. Legal frameworks have been established to address discrimination and promote equal opportunities for all genders. For instance, in Italy, the implementation of gender quotas under Law 120/2011 (also known as Golfo-Mosca) represented a significant milestone, although it was confined to publicly listed and controlled companies. The primary aim of the present study was to conduct a thorough examination of gender equality measures within selected Italian companies, aiming to acknowledge the progress achieved and identify areas for further improvement. Specifically, our focus was on the life sciences sector, because of its high-standing in terms of gender equality and inclusion. An ad hoc questionnaire was developed and distributed to companies within the life sciences sector, including pharmaceutical, medical device, biotechnology, and nutraceutical industries. The questionnaire was completed by human resources professionals. In Italy, the Golfo-Mosca law marked a significant turning point. A comparison will be conducted between 2011 and the current situation to assess whether there has been a spillover effect in companies not directly affected, namely unlisted companies. Compared to 2011, a moderate increase in female representation at the top executive level was observed. None of the companies participating in the survey were directly affected by the Golfo Mosca law; nonetheless, a positive trend might suggest a potential cultural shift towards valuing female talent and leadership.

DELPHI METHOD IN LIFE SCIENCE: MYTH OR REALITY?

Carlotta Galeone[1], Sara Castiglioni[2], Laura Benedan[3], Claudio Pelucchi[4], Paolo Mariani[5]

[1] *University of Milano-Bicocca* (email: laura.benedan@unimib.it)
[2] *University of Milano-Bicocca* (email: sara.castiglioni@unimi.it)
[3] *University of Milano-Bicocca* (email: laura.benedan@unimib.it)
[4] *University of Milano-Bicocca* (email: claudio.pelucchi@unimi.it)
[5] *University of Milano-Bicocca* (email: paolo.mariani@unimib.it)

The Delphi method, originally developed in the mid-20th century, has gained prominence as a structured and iterative technique for achieving consensus among a panel of experts. In the dynamic and multidisciplinary field of life sciences (LS), where diverse perspectives converge, the Delphi method offers a valuable tool for addressing complex research questions, fostering collaboration, and synthesizing expert opinions (1). We aim to elucidate the application of the Delphi method in LS research, emphasizing its efficacy in promoting consensus-building and informed decision-making within the scientific community. Through a review of pertinent literature and illustrative case studies, this presentation will highlight the versatility and adaptability of the Delphi method in addressing the challenges unique to LS research (2). The Delphi method involves a series of structured surveys and iterative rounds of feedback, allowing a panel of experts to converge towards a collective viewpoint. The approach facilitates anonymity, encourages open communication, and accommodates diverse perspectives. By utilizing a systematic process of controlled feedback, the Delphi method enables the identification of areas of agreement and divergence among experts, fostering a more nuanced understanding of complex issues. Case studies from various LS domains will be presented to demonstrate the successful application of the Delphi method. Examples will include its use in establishing disease epidemiology in case of scanty data, prioritizing research objectives, and defining consensus on emerging technologies. The results will showcase the Delphi method as a valuable tool for harnessing the collective intelligence of experts, thereby contributing to evidence-based decision-making in the LS. By critically examining its application, addressing challenges, and presenting successful cases, the presentation will contribute to a deeper understanding of whether the Delphi method is a myth or a tangible reality in advancing consensus-building efforts in this multifaceted field.

References

(1) Humphrey-Murto, Susan, et al. "The delphi method." Academic Medicine 95.1 (2020): 168.
(2) Hasson, Felicity, and Sinead Keeney. "Enhancing rigour in the Delphi technique research." Technological forecasting and social change 78.9 (2011): 1695-1704.

SHORT TIME SERIES IN LABOUR MARKET: A TRAJECTORY ANALYSIS FOR EU COUNTRIES FROM 1995 TO 2022

Paolo Mariani[1], Andrea Marletta[2], Piero Quatto[3]

[1] *University of Milano-Bicocca* (email: paolo.mariani@unimib.it)
[2] *University of Milano-Bicocca* (email: andrea.marletta@unimib.it)
[3] *University of Milano-Bicocca* (email: piero.quatto@unimib.it)

In statistical literature, many contributions investigated the relation between macro-economic variables and the employment rates obtained from the Labour Force Survey. In this work, beyond the total employment rate, other rates referred to particular categories of the population have been analyzed in correspondence with the Gross Domestic Product and an index of inequality income. The source data is Eurostat in the period 1995-2020 for some European countries. The main aim of the work is to detect the presence of an economic growth, measured by GDP, followed by a positive dynamic of the considered employment rates in some EU countries. When this relationship is confirmed, then it is possible to define the concept of inclusive growth. A three-way data analysis approach has been proposed to track the evolution of the relationship during time in each country. The points in the trajectory analysis represent a short time series with not enough observations to achieve good predictions through usual techniques. For this reason, an original approach based on the superior influence of the most recent observations has been used to obtain predictions for the future coordinates of the trajectories. The proposed method also provided prediction intervals in order to display not only the exact next point of the single trajectory but also a measure of the prediction error. The graphical analysis allowed to compare the paths of different countries. The trajectory analysis showed interesting results synchronized to macro-economic events, for example the impact of economic crisis is evident on the trajectories which tend to shrink at that time. About the inclusive growth, only some countries show a clear path toward this direction.

EUROPEAN STATE OF FUTURE INDEX (ESOFI): THE IMPACT OF ECONOMIC, SOCIAL AND ENVIRONMENTAL DIMENSIONS ON THE FUTURE OF THE EUROPEAN UNION

Roberta Di Lorenzo[1], Rocco Mazza[2], Viktoriia Voytsekhovska[3], Rosanna Cataldo[4]

[1] *University of Naples Federico II* (email: dilorenzoroberta1999@libero.it)
[2] *University of Bari Aldo Moro* (email: rocco.mazza@uniba.it)
[3] *Lviv Polytechnic National University* (email: viktoriia.v.voitsekhovska@lpnu.ua)
[4] *University of Naples Federico II* (email: rosanna.cataldo2@unina.it)

Talking about futures in contemporary society is necessary because it makes it possible to plan long-term strategies, promote innovation, and adapt to global changes, thus contributing to sustainable growth and collective well-being. In recent years, the European Union, facing a complex series of economic, social, conflict and health crises, has shown a marked awareness of the significant gaps in its governance system and the need to think about forward-looking strategies for the future. In response to this need, in 2021, the European Commission organized the first Conference on the Future of the European Union, an initiative shaped by the need to critically address the shortcomings that emerged during the crises and to actively engage citizens in the debate on the future prospects and developments of European integration, thereby reframing the common path toward a stronger and more cohesive Union. So, the goal of this contribution is to construct a composite indicator that can provide a clear and comprehensive perspective on the future of European nations, serving as a strategic guide for policy formulation attentive to emerging challenges and changing dynamics. The starting point of the present research was to analyze the discussions and proposals of the nine key themes derived from the Future of the European Union Conference, and based on this conceptual framework, indicators were selected, creating a dataset consisting of 23 indicators. Through the application of the Partial Least Squares Path Modeling (PLS-PM) statistical technique, the synthetic indicator "European State Of Future Index" (ESOFI) was created, an advanced model useful for formulating targeted strategies and policies. The results highlight the EU's commitment to climate change policies; however, it emphasizes the importance of not neglecting key aspects such as health, youth and security to ensure an overall better and sustainable future.

RESEARCHING DISCOURSES ON EMERGING TECHNOLOGIES: HOW INTEGRATING QUALITATIVE AND QUANTITATIVE DATA CAN IMPROVE THE QUALITY OF SOCIAL MEDIA DATA

Francesco Amato[1], Biagio Aragona[2], Dario Chianese[3], Mattia De Angelis[4]

[1] *University of Naples Federico II* (email: francesco.amato@unina.it)
[2] *University of Naples Federico II* (email: biagio.aragona@unina.it)
[3] *University of Naples Federico II* (email: dario.chianese@unina.it)
[4] *University of Naples Federico II* (email: mattia.deangelis@unina.it)

High-performance computing (HPC) is an enabling technology that guarantees high computing capabilities for heterogeneous applications. These systems solve large-scale computational problems requiring considerable computation, such as climate simulation, weather prediction, pharmaceutical research, aircraft design, financial market analysis, and more.

With this contribution, we want to explore how the results obtained on digital tools differ from those integrated with the knowledge of key informants. How integrating qualitative data can help gain a greater understanding of social media data.

To do this, we collected data by different means. The first focused on the statistical analysis of X's posts on the topic of HPC. The second involved the integration of a qualitative phase with a quantitative one.

We interviewed 25 experts in HPC technologies and applications. A content analysis was carried out after the transcription of the interviews, and keywords were extracted from the Google Search and Trends platform. From Google services, we extracted the associated queries and topics. We extracted the digital traces from the social media X by integrating these new keywords.

The integration of the extractions on X and Google served to compare the results obtained by digital tools with those of the interviews and see how the perspectives on HPC may differ.

The results of the two approaches highlighted the contribution of qualitative data in obtaining greater understanding and meaning from social media data. This enriches the research conducted on X and the interpretation of the results.

Integrating different data types is effective in studying the discourses on emerging technologies whose knowledge on social media may be fragmented.

DATA JOURNALISM IN THE FIGHT AGAINST DISINFORMATION

Luigi Rossetti

University of Campania Luigi Vanvitelli (email: luigi.rossetti1@studenti.unicampania.it)

The rise of large-scale AI and machine learning models has sparked an explosion in falsified information, marked by hallucinated or synthetic contents (text, images, videos, audio) crafted through digital technologies. According to the World Economic Forum's Global Risks Report for 2024, this phenomenon is a significant short-term threat necessitating urgent responses. Data journalism emerges at this critical juncture, bringing together data with communication to highlight a human centric design approach. This paper provides an overview of data journalism, including how data visualization and storytelling can be applied into insights for effective communication. The goal is to help in creating a data driven culture, transforming data into visually shared stories for a clearer communication and to increase ethical transparency in the representation of facts. Big data and journalism combat disinformation through sophisticated data analysis and epistemological considerations (1). A mixed-methods approach that includes case studies, content analysis, data and AI ethics is required. Anthropocentric storytelling, when combined with strategic data analysis tools such as data scraping, mining, statistical analysis, and visualization, emerges as a potent strategy to combat disinformation (2). However, there is a trade-off between using narratives for persuasive clarity and the risk of manipulating information (3). This dichotomy underscores the importance of ethical storytelling in data journalism. The challenges of data journalism especially in the context of AI-Ethics, are complex. Key recommendations involve improving data literacy, using verification of multiple sources and knowledge analysis to ensure ethical alignment with AI practices.

References

(1) Lewis, S. C., & Westlund, O. (2015). Big data and journalism: Epistemology, expertise, economics, and ethics. Digital journalism, 3(3), 447-466

(2) Lee, B., Riche, N. H., Isenberg, P., & Carpendale, S. (2015). More than telling a story: Transforming data into visually shared stories. IEEE computer graphics and applications, 35(5), 84-90.

(3) Krause, R. J., & Rucker, D. D. (2020). Strategic storytelling: When narratives help versus hurt the persuasive power of facts. Personality and Social Psychology Bulletin, 46(2), 216-227.

DIGITAL DATA AND WELFARE POLICIES: FROM INFORMATION SYSTEMS TO NEW INNOVATIONS OF PARTICIPATORY DATA GOVERNANCE

Luca de Luca Picione[1], Lucia Fortini[1], Domenico Trezza[3]

[1] *University of Naples Federico II* (email: giuseppeluca.delucapicione@unina.it)
[2] *University of Naples Federico II*
[3] *University of Naples Federico II* (email: domenico.trezza@unina.it)

The digital revolution is greatly shaping the world of educational and social services. Not only does it change risks, bringing forth entirely new social needs, but it also alters the tools to counteract them, such as the way social policy is conducted. The digitalization of social data certainly facilitates the processes of identification and response to these needs, but on the other hand, it also brings forth new challenges. Among these challenges, certainly, is the issue related to data governance. While in the past, attention was focused on consolidating the IT infrastructures of social systems and creating socio-normative devices to ensure public access to data (Mauri, 2007; Rinaldi, 2012), today the evolution of artificial intelligence systems for data management presents us with many opportunities and new critical points to address related to ethics, possible algorithmic biases, and data transparency (Amaturo and Aragona, 2019; Vesan and Campedelli, 2023).
In light of these premises, the paper presents the experience of data governance in social policies of Campania Region, in relation to two specific cases: one related to the Social Information System, that is the IT structure for service planning and thus social data management. The other case concerns the institutional experimentation of Govern-AI (*GOVERNance assistance for social areas by Artificial Intelligence*), a research and intervention program aimed at AI-based assistance in local welfare governance. The approach used is that of constructing social digital data through a shared and transparent process.

IMPACT OF THE RUSSIAN INVASION OF UKRAINE ON COAL MARKETS: EVIDENCE FROM AN EVENT-STUDY APPROACH

Paola Cerchiello

University of Pavia (email: paola.cerchiello@unipv.it)

The paper examines the immediate market response of the coal-related industry to the onset of the Russian full-scale invasion of Ukraine. The findings indicate a significant and negative reaction in the European region on the 21st and 24th of February, followed by an increase in average abnormal returns (AARs) on the first day following the invasion. As for Emerging countries, no results were observed for China, while India registered positive AARs on the second and third day of the war.
Over the designated 14-day event period, positive Cumulative Average Abnormal Returns (CAARs) are evident in all examined sub-samples, though statistical significance is observed only within the Titans subsample.

STATISTICAL LEARNING METHODS FOR EARLY DETECTION OF CORPORATE CRISES

Donato Riccio[1], Giuseppe Maria Bifulco[2], Francesco Paolone[3], Andrea Mazzitelli[4], Fabrizio Maturo[5]

[1] University of Campania Luigi Vanvitelli (email: donato.riccio@studenti.unicampania.it)
[2] University of Naples Federico II (email: giuseppe.bifulco@unina.it)
[3] Universitas Mercatorum (email:francesco.paolone@unimercatorum.it)
[4] Universitas Mercatorum (email:andrea.mazzitelli@unimercatorum.it)
[5] *Universitas Mercatorum* (email: fabrizio.maturo@unimercatorum.it)

First forecasting models to diagnose the state of corporate health date back to the 1960s and 1970s. These were entirely based on balance sheet indexes, which still represent a valid tool for preventing the company's state of health. The most used in the literature were certainly profitability (ROE, ROI, ROS, and Turnover), liquidity, and capital-financial solidity (e.g.debt ratio) (1,2,3). In recent years, the economy has gone through a deep structural crisis that has sanctioned the settling of income, cash and capital levels significantly lower than in the past. In many cases, this crisis has led to insolvency and bankruptcy. There is, therefore, an urgent need to identify a series of warning signs to activate a recovery process promptly and effectively before the prospects of business continuity are compromised.
To this end, there is a growing need to combine quantitative models based on advanced statistical learning techniques that consider large volumes of data, the temporal aspect of the balance sheet data, the dynamic aspect of the context variables and the interactions of these data over time. We propose using ensemble techniques because they overcome many issues of classical parametric models. The study shows that using tree-based statistical learning methods with optimization of the model's hyper-parameters provides very high results in terms of accuracy and can automatically consider all possible interactions. The latter aspect allows us to discover relevant features never considered in past studies. We often have large datasets regarding sample size and the number of variables in the business field. Our study shows that integrating the latest statistical learning techniques in this area significantly benefits prediction and explainability. Further studies are recommended to introduce context variables and spatial analysis since these techniques have been little used in the business field to predict the state of crisis.

References

(1) Altman EI. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance. 1968;23(4):589-609.
(2) Jones S, Hensher DA. Predicting firm financial distress: A mixed logit model. Journal of Accounting and Public Policy. 2004;23(6):467-487.
(3) Shumway T. Forecasting bankruptcy more accurately: A simple hazard model. Journal of Business. 2001;74(1):101-124.

ASSESSING POLLUTION'S IMPACT ON QUALITY OF LIFE WITH FUNCTIONAL REGRESSION

Gianmarco Borrata[1], Antonio Balzanella[2], Rosanna Verde[3]

[1] *University of Naples Federico II*
[2] *University of Campania Luigi Vanvitelli* (email: antonio.balzanella@unicampania.it)
[3] *University of Campania Luigi Vanvitelli* (email: rosanna.verde@unicampania.it)

In this paper we present a new regression method for data represented in the form of distributions (1). Unlike traditional functional regression methods, our contribution lies in a specific data transformation. We employ a logarithmic transformation on the derivative quantile functions linked to the distributional data. Let E be a set of N objects observed on p distributional variables {X1, X2, . . ., Xp}. Each object is represented by p probability density functions (pdf 's), or empirical ones, denoted fij(s). In consideration of the most recent developments in Distributional Data Analysis (DDA), we introduce a transformation of the quantile functions (qf 's), associated to the (pdf 's), named Logarithm Derivative Quantile (LDQ) functions lij(t) defined in the interval [0, 1]. A similar transformation was introduced in (2) to map density probability functions in a Hilbert space. This distributional processing of the data has the advantage of allowing an analysis of the new functions and being able to return from the achieved results to the original quantile functions, through an inverse transformation. We consider an extension for distributional data on the LDQ functions by using a functional data representation. For each distributional variable Xj (for j=1, . . ., p), we assume that the LDQ functions are represented like functional data (3) by considering a B-splines smoothing in the points corresponding to the quantile of the distributions. In the context of applying the regression method for distributional data and the transformation of quantile functions (LDQ), our objective is to utilize this new regression method to thoroughly evaluate the influence of environmental pollution on a quality of life.

References

(1) Bock H., Diday E.: Analysis of symbolic data: exploratory methods for extracting statistical information from complex data. Springer Science & Business Media, (1999).
(2) Petersen A., Muller H.: Functional data analysis for density functions by transformation to a Hilbert space. The Annals of Statistics, Ann. Statist. 44(1), 183-218, (2016).
(3) Ramsay, J. O. & Silverman, B. W. Functional Data Analysis, 2nd Edition, Springer, New York. (2005)

LAPLACIAN EMBEDDING AND SPECTRAL CLUSTERING FOR THREE-WAY DATA

Cinzia Di Nuzzo[1], Salvatore Ingrassia[2]

[1] *University of Catania* (email: cinzia.dinuzzo@unict.it)
[2] *University of Catania* (email: salvatore.ingrassia@unict.it)

A new spectral clustering method for three-way data is introduced. Spectral clustering is a two-step sequential method that involves the dimensionality reduction of data through what is referred to as *Laplacian embedding*; the second step entails applying a clustering algorithm (usually the *k-means* algorithm, but also mixture models have been taken into account) to partition the data into K clusters. While spectral clustering methods have been traditionally applied to two-way data, recently, we proposed a spectral clustering approach in three-way data in (1) and (2). The goal is to partition a three-way dataset composed by N units, M variables and H occasions into K clusters. The novelty of the method here proposed concerns modelling a set of H Laplacian matrices $\mathbf{L}_h$ (for $h = 1, . . ,H$) of size $N \times N$ as follows:

$$\mathrm{Lh}=\mathrm{AChA}'+\mathrm{Eh} \text{ such that } \mathrm{AA}'=\mathrm{I} \text{ for h=1,….,H.} \qquad (1)$$

$\mathbf{L}_h$ is the Laplacian matrix, $\boldsymbol{A}$ is the $N \times K$ matrix representing Laplacian embedding, and $\mathbf{C}_h$ is a diagonal $K \times K$ matrix with non-negative elements.
A least squares approach is considered to estimate the model, based on the following minimization

$$\mathrm{h}=1\mathrm{H}||\mathrm{Lh}-\mathrm{ALhB}'||2, \qquad (2)$$

where $\boldsymbol{B}$ is an $N \times K$ matrix introduced for algorithmic reasons. In this context, a significant innovation comes from results given in [3] and the related algorithm known as *Indort*; in particular in (3) is proved that if the matrices $\boldsymbol{L}_{\#}$ are positive definite and the elements of $\mathbf{C}_h$ are non-negative, then the resulting matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ in (2) from the Indort algorithm are equal.
These results allow obtaining a unique configuration of Laplacian embedding common to all H Laplacian matrices $\boldsymbol{L}_{\#}$, making the matrix $\boldsymbol{A}$ representative for the entire dataset. Once the Laplacian embedding is computed, data can be clustered according to the standard spectral approaches.

References

(1) Di Nuzzo C., Ingrassia S. (2023). Three-way Spectral Clustering, in ``Brito P., Dias J.G., Lausen B., Montanari A., Nugent R. (Eds.) *Classification and Data Science in the Digital Age*", Studies in Classification, Data Analysis, and Knowledge Organization, Springer, 111-118.

(2) Di Nuzzo, C., Ingrassia, S., Vicari, D. (2022). An INDSCAL-Type Approach for Three-Way Spectral Clustering. In ``García-Escudero L.A., Gordaliza A., Mayo-Iscar A., Lubiano Gomez M.A., Gil M.A, Grzegorzewski P., Hryniewicz O. (Eds.) *Building Bridges between Soft and Statistical Methodologies for Data Science*", Advances in Intelligent Systems and Computing, Springer, **1433**, 128-135.

(3) Dosse, M.B., ten Berge, J.M. & Tendeiro, J.N. (2011). Some New Results on Orthogonally Constrained Candecomp. *Journal of Classification,* **28**(2), 144–155.

RANDOMLY PERTURBED RANDOM FORESTS

Angela Montanari[1], Laura Anderlucci[2]

[1] *University of Bologna* (email: angela.montanari@unibo.it)
[2] *University of Bologna* (email: laura.anderlucci@unibo.it)

In supervised classification, a change in the distribution of a single feature, a combination of features, or the class boundaries, may be observed between the training and the test set. This situation is known as dataset shift (1). As a result, in real data applications, the common assumption that the training and testing data follow the same distribution is often violated. For example, when dealing with CUP (Cancer of Unknown Primary) samples, the problem is to correctly identify the (unknown) starting point of metastatic cancer cells; however, when cancer spreads, the secondary cancer cells may look like abnormal versions of the primary cancer cells and classifiers trained on the latter ones are no longer accurate.
Dataset shift might be due to several reasons; the focus is on what is called "concept shift", namely the conditional probability of the features (X) given the response (Y) differs from training to test set. The aim is to address dataset shift in supervised classification problems.
In order to address dataset shift we propose to randomly introduce more variability in the training set by sketching the input data matrix resorting to random projections (2) of units. We then modify the random forests algorithm to involve sketched, rather than bootstrapped, versions of the original data.
Results on real data show that perturbing the training data via matrix sketching (3) improves the prediction accuracy of test units that have a different distribution in terms of variance structure.
The application of the proposed methodology to the identification of the unknown starting point of metastatic cancer cells shows that data perturbation based on sketching produces promising results; however, it is important to understand which group characteristics benefit from the projection via random subspaces.

References

(1) J.G. Moreno-Torres, et al. (2012) A unifying view on dataset shift in classification. Pattern recognition 45,1, 521–530.
(2) D. C. Ahfock, W.J. Astle, S. Richardson (2020), Statistical properties of sketching algorithms, Biometrika, 102, 2, 283–297.
(3) R. Falcone, L. Anderlucci, A. Montanari, (2022). Matrix sketching for supervised classification with imbalanced classes, Data Mining and Knowledge Discovery, 36, 174-208.

PROBABILISTIC DISTANCE CLUSTERING FOR MIXED-TYPE DATA TO ANALYZE STUDENT DATA

Francesco Palumbo[1], Cristina Tortora[2]

[1] *University of Naples Federico II* (email:fpalumbo@unina.it)
[2] *San José State University* (email: cristina.tortora@sjsu.edu)

The tremendous rise in data complexity in recent years has led to the need for new statistical methods. Nowadays, data are recorded and stored with purposes that differ from statistical analysis and often consist of mixed-type data that cannot be treated under any specific distributional assumptions. Therefore, mixed-type data analysis has attracted the interest of many from the clustering and classification domain (1). When dealing with mixed data types, with many variables and a big number of units, geometric partitioning clustering algorithms remain more appealing and likely the most broadly used. Within this large family, this work considers probabilistic distance (PD) clustering (2), which is a distribution-free, probabilistic clustering algorithm that can work with any general distance. PD clustering for mixed-type data measures the homogeneity among the units using Gower's distance and assigns the units to the given K clusters by solving a numerical optimization problem. Simulation studies prove the method's efficacy under several diverse conditions [3].
This proposal presents a PD clustering application on a real mixed-type data set that was recorded on a group of students enrolled in the first year of an ungraduated program in psychology. The final aim is to identify homogenous groups of students according to their mathematical prerequisite knowledge, socioeconomic characteristics, and some psychological traits that can affect academic performance.

References

(1) Van de Velden M, Iodice D'Enza A, Markos A. Distance-based clustering of mixed data. Wiley Interdisciplinary Reviews: Computational Statistics. 2019 May;11(3):e1456.
(2) Ben-Israel A, Iyigun C. Probabilistic D-clustering. Journal of Classification. 2008 Jun;25:5-26. [3] Tortora C, Palumbo F. Clustering mixed-type data using a probabilistic distance algorithm. Applied Soft Computing. 2022 Nov 1;130:109704.

DIFFERENTIAL ERROR EFFECTS ON CLUSTERING MIXED-TYPE DATA

Valentina Veronesi[1], Marianthi Markatou[2]

[1] *University at Buffalo* (email: vverones@buffalo.edu)
[2] *University at Buffalo* (email: markatou@buffalo.edu)

Clustering serves as a fundamental tool in various scientific disciplines, offering a method to analyse a wide spectrum of data types. Especially in disciplines collecting data from human sources, the most common data type is the mixed one, comprising a combination of interval scale, nominal, and ordinal variables. Applying a traditional clustering algorithm carries the implicit and overly idealistic assumption of error-free data, which is rarely the reality in data collection and processing. This is especially true when humans are involved in some parts of the collection process. Transposition errors, for which numerical values of variables are swapped (a common phenomenon in clinical fields, where data are manually recorded), and recall bias, i.e., the inaccurate or incomplete recollection of past events, are just two examples that underscore the prevalence of data inaccuracies in real-world datasets. The influence of observational errors on clustering algorithms have the potential to impact the quality of the results. While efforts have been made to mitigate this influence in the context of interval scale data, the consequences of using already available clustering methods not specifically developed for handling errors are not known. We concentrate on the effects of differential measurement errors and misclassification on clustering. Our study compares the robustness of five clustering algorithms for mixed type data, namely k-prototypes, Modha-Spangler, KAMILA, HyDaP, and PDQ, in the presence of measurement error and misclassification (MEM), through both a Monte Carlo study and a real world application. This study provides comprehensive guidelines for users to align clustering algorithm selection with data characteristics and MEM. Additionally, knowing that the analysis of errors in regression has provided valuable insights into understanding the implications of errors, we provide equivalent definitions, given the lack of an analogous theory in cluster analysis.

THE CTA-PLS APPROACH FOR SEM: APPLICATIONS

Mario Angelelli[1], Enrico Ciavolino[2]

[1] *University of Salento* (email: mario.angelelli@unisalento.it)
[2] *University of Salento* (email: enrico.ciavolino@unisalento.it)

The distinction between the formative and reflective measurement modes in Structural Equation Modelling (SEM) has both methodological and practical implications. Measurement misspecification could lead to different interpretations of the causal structure (causal vs. effect indicators) (1) and biased estimates even in the structural model. This contribution focuses on the need for new methodologies that adapt to different experimental settings in psychometric assessments and evaluations and guide the adoption of an adequate measurement mode. Our investigation relies on algebraic constraints (tetrad conditions) arising for the observed covariance matrix under a reflective measurement mode (1). Applications are often subject to limited data sample sizes and deviations from distributional assumptions (e.g., normality), which requires robust procedures that do not rely on asymptotic theory, especially for explorative application studies. In line with our research question, we concentrate on CTA-PLS, which is a combination of non-parametric estimation approaches (Partial Least Squares) with Confirmatory Tetrad Analysis (2). CTA-PLS does not produce a unique statistical test based on assumptions underlying asymptotic theory; instead, it produces a test for each algebraic relation (tetrad), which leads to a multiple hypothesis-testing framework. In this work, we will show how the CTA-PLS approach is specified in different case studies, discussing sources of uncertainty and comparing different adjustment criteria to highlight their performance and limitations in experimental research. We will discuss the comparison of CTA-PLS with other measures to assess the quality of the measurement model, pointing out the complementary information they provide. A sample-based approach is suggested to investigate the test power using different adjustment criteria starting from the observed data, without *a priori* ground truth. The presented case studies will address different experimental conditions based on sample size, number of manifest variables, and number of latent variables.

References

(1) Bollen, K. A., & Ting, K. F. (2000). A tetrad test for causal indicators. *Psychological methods*, *5*(1), 3.

THE CTA-PLS APPROACH FOR SEM: SIMULATION

Mattia Cefis[1], Maurizio Carpita[2]

[1] University of Brescia (email: mattia.cefis@unibs.it)
[2] University of Brescia (email: maurizio.carpita@unibs.it)

Structural Equation Modelling (SEM) analyses latent constructs through observed indicators, defining connections between latent and manifest variables as reflective or formative. Reflective measurement assumes latent constructs exist independently of indicators, while formative measurement sees the latent variable as a composite of observed manifest variables. Partial Least Squares SEM (PLS-SEM) efficiently handles both measurement modes, yet selecting between them poses computational and conceptual challenges. This contribute pays attention on Confirmatory Tetrad Analysis in PLS-SEM (CTA-PLS, (1)), a multiple test that extends the tetrad test to PLS-SEM and examines it from a decision
making standpoint. The choice between reflective and formative modes, formalized through assumptions about vanishing tetrads, relies on multiple hypothesis testing. To aid PLS-SEM researchers and practitioners, a simulation study is presented. Through a data generation process, reflective and formative measurement models are rigorously tested using CTA-PLS, evaluating its actual significance level and power across various sample sizes and manifest variable scenarios. The simulation study compares the performance of different criteria for adjusting the statistical significance level, including the well-established Bonferroni correction, and alternatives like the Benjamini-Hochberg and Benjamini-Yekutieli corrections. The findings highlight the superior performance of the Benjamini Hochberg method, particularly in terms of test power, especially when dealing with small sample sizes and numerous manifest variables. This contribute offers valuable insights to guide researchers and practitioners in making informed decisions about measurement modes in PLS-SEM, enhancing the robustness and reliability of their analyses.

ON THE SELECTION OF A UNIDIMENSIONAL SET OF ITEMS

Alessio Farcomeni

University of Rome Tor Vergata (email: alessio.farcomeni@uniroma2.it)

Assessment of dimensionality of a latent trait, and estimation of multidimensional models, is now well studied in the psychometric literature. In this work we take a different perspective, and outline a methodology to select, among a set of candidate items, a subset that share a single common latent trait and is as large as possible. We build on a specific definition of multidimensional latent model that is based on a discrete latent variable. We then implement a simple stochastic search strategy, which is computationally feasible despite the gigantic number of possible subsets. Our strategy involves sequentially testing for unidimensionality versus a local alternative, until a stopping rule is satisfied. At convergence, the final subset is verified using the much more stringent condition that is should pass a test for unidimensionality against all possible bidimensional alternatives. We illustrate the approach by constructing a European measurement scale for material deprivation, with an initial candidate set of twenty-three items and a final set of eight items that can be argued to be unidimensional.

A SECOND-ORDER PATH MODELLING APPROACH FOR THE ASSESSMENT OF HONEY BEE COLONY HEALTH

Anna Simonetto[1], Gianni Gilioli[2]

[1] University of Brescia (email: anna.simonetto@unibs.it)
[2] Gianni Gilioli, University of Brescia (email: gianni.gilioli@unibs.it)

Honeybees play a crucial role in providing ecosystem services like food (e.g., honey, beebread, and royal jelly) and pollination essential for maintaining biodiversity and food security. The provision of these services is heavily influenced by the health status of the bees. The health status of a colony of bees is highly impacted by heterogeneous factors including environmental conditions, chemical and biological stressors, beekeeper conditions and practices, and agricultural practices. This complexity makes it challenging to implement management strategies that both preserve honeybee health and ensure productivity and economic viability of beekeeping. Honeybee health, like human health, cannot be directly measured. The conceptual model of Health Status Index (HSI) was proposed by EFSA (1), but quantitative tools for its operationalization have not been presented. In this study, we applied a second-order structural equation model (SEM) within a Partial Least Squares Path Modeling (PLS-PM) approach to estimate the HSI. Collaborating with experts, we identified six dimensions of colony health as the primary level of the model. Honeybee health is defined as a second-order construct (2). Furthermore, the model includes the estimation of three latent dimensions related to external drivers (environmental conditions and beekeeping practices) and their relationship with the six health dimensions. The model has been applied to a simulated data set representing seven realistic scenarios related to the external drivers and one of the six health dimensions (disease and infection). The probability distributions of parameters defining the scenarios were estimated through expert opinion. The model was able to discriminate the health status of the honeybee colonies according to the characteristics of each scenario. The HSI provided a good capacity to integrate different types of data.

References

(1) EFSA Panel on Animal Health and Welfare (AHAW). (2016). Assessing the health status of managed honeybee colonies (HEALTHY-B): a toolbox to facilitate harmonised data collection. EFSA Journal, 14(10), e04578.
(2) Sanchez, G., 2013. PLS path modeling with R. Trowchez Editions, Berkeley, pp. 383.

INNOVATIVE INCLUSIVE KITCHEN SYSTEM FOR VISUALLY IMPAIRED: AIRFLOW-INDUCED HEAT NOTIFICATION FOR SAFE COOKING

Francesco Paolicelli[1], Antonio Carmelo Di Gioia[2], Vito Santarcangelo[3], Davide Scintu[4]

[1] *Thegg Domotica Srl*
[2] *Thegg Domotica Srl*
[3] *iInformatica Srl* (email: vito@iinformatica.it)
[3] *iInformatica Srl*

This work presents a novel inclusive kitchen system designed to revolutionize cooking safety and accessibility for visually impaired individuals. At the heart of this system lies an innovative airflow feedback mechanism. This system crucially provides information about the operational status and temperature of cooking surfaces, particularly stovetops, using distinct airflow patterns.
These patterns are intelligently configured based on a specific semantic knowledge base, which tailors the system's responses to enhance usability and precision. Such a design is essential for visually impaired users, who cannot depend on traditional visual cues to ascertain whether a stove is active or hot. The system functions by emitting controlled air jets, which vary in intensity to indicate the stove's temperature. A gentle airflow signals a cool or inactive stove, while a more intense airflow denotes a hot surface. This tactile and thermal feedback mechanism is extended to the stove's control panel as well, where airflow variations communicate different functional statuses like heat intensity and operational state. This dual application of the airflow system ensures that users can both operate the stove safely and be aware of its current state. The inclusive kitchen system is developed with the key objectives of the United Nations 2030 Agenda for sustainable development in mind, aiming to promote the well-being and autonomy of people with disabilities. By integrating advanced sensory feedback technologies, this system marks a significant advancement in creating inclusive home environments.

References

(1) Kim, M., Hwang, S., Choi, K., Oh, Y., Lim, D. (2022). Vision-Based Cooking Assistance System for Visually Impaired People. In: Stephanidis, C., Antona, M., Ntoa, S. (eds) *HCI International 2022 Posters. HCII 2022. Communications in Computer and Information Science*, vol 1580. Springer, pp 540–547.

ANALYSIS BY ARTIFICIAL INTELLIGENCE OF THE BRAND PROMISE OF A CHROMOGENIC LABEL

Francesco Vena[1], Leonardo Vena[2], Letizia Vena[3], Francesco Giannone[4], Saverio Gianluca Crisafulli[5], Emilio Massa[6], Vito Santarcangelo[7], Marco Vito Calciano[8]

[1]*Lucano 1894 Srl*
[2] *Lucano 1894 Srl*
[3] *Lucano 1894 Srl*
[4] *Lucano 1894 Srl*
[5] *iInformatica Srl*
[6] *iInformatica Srl*
[7] *iInformatica Srl*
[8] *Zio Startup Srl*

The present research work concerns the presentation of the initiative of a chromogenic label applied to a bottle of Amaro Lucano in order to suggest the right consumption linked to an ideal temperature. Thanks to a chromogenic layer composed of thermochromic materials, i.e., transparent substances subject to reversible modification of optical properties depending on temperature through a chemical reaction, perfectly integrated in the layout of a label, it is possible to detect with the appearance of the blue color the perfect temperature to consume Amaro Lucano. The objective of this research work is to evaluate how such an initiative (in terms of generative storytelling) is perceived by an "alien," a hypothetical focus group of users who do not know Amaro Lucano or who know it through OSINT sources, exploiting the potential of explainable artificial intelligence (XAI). Such an innovative approach called "NIGG" (Names and Images -> Items' Graph -> Generative Artificial Intelligence) enables brand promise prediction by estimating how an initiative may affect consumer perceptions. This paper discusses such innovative and patented experimental approach for brand promise analysis, some results and the wide scope of application.

References

(1) Hakami, A., Srinivasan, S. S., Biswas, P. K., Krishnegowda, A., Wallen, S. L., & Stefanakos, E. K. (2022). Review on thermochromic materials: development, characterization, and applications. Journal of Coatings Technology and Research, 19(2), pp. 377-402.

INNOVATIVE DYNAMIC SYSTEM FOR INCLUSIVE SCREENING IN OPTICS AND ORTHOPTICS USING TAILOR-MADE 3D-PRINTED DIAGNOSTIC GOGGLES AND AI-DRIVEN MODELLING AND EXERCISES

Giuseppe Scavone[1], Angelo Romano[2], Vito Santarcangelo[3], Massimiliano Giacalone[4]

[1] *Centro Rham Srl*
[2] *iInformatica Srl*
[3] *iInformatica Srl* (email: vito@iinformatica.it)
[4] *University of Campania* (email: massimiliano.giacalone@unicampania.it)

This work introduces an innovative system designed to revolutionize the field of optics and orthoptics health screening. At the core of this system is a set of tailor-made diagnostic goggles, created using advanced 3D printing technology, and by a sophisticated expert system based on artificial intelligence (AI) module, which precisely tailors the 3D model according to the specific screening requirements and patient's specific physical characteristics. A significant aspect of this system is its integration with a tablet, which, when paired with the goggles, delivers customized visual exercises. These exercises, including attention to visual details, tracking, and chromatic problem detection are dynamically generated by the expert system. The system's intelligence lies in its ability to adapt to the user's specific needs, taking into consideration the type of goggles used, user information, and the patient's specific screening objectives.

Moreover, the system intelligently tracks the user's past interactions. This information is leveraged to continually generate novel scenarios, thus preventing users from relying solely on memory from previous tasks and ensuring consistently engaging and effective screening sessions.

This innovative system aligns with the United Nations 2030 Agenda for sustainable development, contributing to the advancement of global health and well-being. By harnessing the power of AI and 3D printing, it offers a novel, inclusive approach to optical and orthoptic health screening, emphasizing personalized care and continuous adaptation to individual needs.

References

(1) König, H.-H. Barry, J.-C. Leidl, R. Zrenner, E. (2002). Economic Evaluation of Orthoptic Screening: Results of a Field Study in 121 German Kindergartens. *Invest. Ophthalmol. Vis. Sci.* **43**(10), pp. 3209-3215.

(2) Sanchez, I. Ortiz-Toquero, S. Martin, R. de Juan, V. (2016) Advantages, limitations, and diagnostic accuracy of photoscreeners in early detection of amblyopia: a review. *Clinical Ophthalmology*, **10**, pp. 1365-1373.

(3) Donaldson, L.A., Karas, M.P., Charles, A.E. Adams, G.G.W. (2002), Paediatric community vision screening with combined optometric and orthoptic care: a 64-month review. *Ophthalmic and Physiological Optics*. **22**, pp. 26-31.

(4) Anker, S. Atkinson, J. Braddick, O. Ehrlich, D. Hartley, T. Nardini, M. Wade, J. (2003). Identification of Infants with Significant Refractive Error and Strabismus in a Population Screening Program using Noncycloplegic Videorefraction and Orthoptic Examination. *Invest Ophthalmol. Vis. Sci.* **44**(2), pp. 497-504.

AN INNOVATIVE APPROACH FOR THE INTERCONNECTION AND MONITORING OF A TABLE SOCCER

Nicola Favale[1], Alessandro D'Alcantara[2], Vito Santarcangelo[3], Marco Calciano[4], Massimiliano Giacalone[5]

[1] *Different Game Srl*
[2] *iInformatica Srl*
[3] *iInformatica Srl* (email: vito@iinformatica.it)ù
[4] *Zio Starup Srl*
[5] *University of Campania* (email: massimiliano.giacalone@unicampania.it)

Different Var is an experimental intelligent system that allows through application and sensors to interconnect table soccer tables for the purpose of real-time monitoring of scores (in relation to detected goals) during games directly from a convenient browser-based intranet dashboard. Each football table also features its own camera that allows the dynamics of the game to be filmed and allows replays to be taken when the goal is detected or via appropriate trigger signal. In this way, a video book is created for each game, allowing memories of convivial moments or competitions over time. The table also features an internal UV-C LED system that also allows sanitization of the game balls. The system by correlating sensor data with spectator's perceptions data, the fair play analysis and the computer vision module applied to the videos enables an innovative data analysis approach of individual players' performances. The system is protected by patent for industrial invention filed No. 102023000013596 and its design is protected by registered design.

References

(1) T. Weigel et al. (2004). Adaptive Vision for Playing Table Soccer, KI 2004: Advances in Artificial Intelligence
(2) R. Janssen (2009). Real-Time Ball Tracking in a Semi-automated Foosball Table, RoboCup 2009

UNVEILING CORRUPTION RISKS IN PUBLIC PROCUREMENT: THE IMPERATIVE OF BIG DATA MANAGEMENT

Michela Gnaldi[1], Simone Del Sarto[2]

[1] *University of Perugia* (email:michela.gnaldi@unipg.it)
[2] *University of Perugia* (email:simone.delsarto@unipg.it)

In the contemporary landscape of public governance, the management of corruption risks in public procurement processes stands as a critical challenge that demands innovative solutions. As governments strive for increased transparency, accountability, and efficiency, big data analytics emerges as a powerful ally in the fight against corruption. The immense volume of data generated throughout the procurement lifecycle provides a treasure trove of insights, enabling policymakers, auditors, and anti-corruption agencies to identify irregularities, enhance oversight, and safeguard public resources. However, harnessing the full potential of big data in combating corruption is not without its hurdles. This article delves into the pressing necessity of utilizing big data analytics in public procurement, shedding light on the complexities surrounding data quality, accuracy, and coverage. While the promise of big data lies in its capacity to unveil patterns, detect anomalies, and facilitate evidence-based decision-making, the success of these endeavours is intrinsically tied to the reliability of the underlying data. Inconsistencies, inaccuracies, and incompleteness within most European and national datasets can compromise the integrity of analyses and, consequently, the effectiveness of corruption detection mechanisms. By referring to the recently ended European project CO.R.E. (Corruption Risk Indicators in Emergency - Grant agreement n. 101038790), in this contribution we illustrate the huge potentials of big data in the context of combating corruption in public procurement, examining the pivotal role it plays, while also scrutinizing the inherent challenges related to data quality and coverage and paving the way for more robust strategies and methodologies to harness big data's potential.

BIG DATA EVOLUTION IN THE ITALIAN JUDICIAL FRAMEWORK

Gianfranco Piscopo[1], Vincenzo Basile[2], Maria Longobardi[3], Massimiliano Giacalone[4]

[1] *University of Naples Federico II* (email: gianfranco.piscopo@unina.it)
[2] *University of Naples Federico II* (email: vincenzo.basile2@unina.it)
[3] *University of Naples Federico II* (email: maria.longobardi @unina.it)
[4] *University of Campania Luigi Vanvitelli* (email: massimiliano.giacalone@unicampania.it)

The Big Data Analysis allows the management of large amounts of data of different nature and from various sources, having great importance also in judicial proceedings. In the investigative and judicial field, analysis of correlations, semantic enrichment and sentiment analysis have proved valuable tools for feedback of projective type: such a statistical analysis thus appears to be of fundamental support in judicial activity (1). Moreover, in Italy there have been massive investments in technological innovation with the promise of transforming justice into a quality service, increasing its ability to act in an effective, efficient, transparent way, and in line with the actual citizens' expectations, allowing them to file a lawsuit and to forward it to the competent court, in an automated way. The process of technological innovation in justice is likely to produce items incompatible with a background that was designed for the manuscript, where the requests have been just applied to what was there in the past (2).
This paper aims to describe the main statistical tools for innovative problem-solving of the jurists' activities, defining a horizon of application ranging from the study of regulatory and jurisprudential corpora through information extraction techniques, network analysis, application of complexity models, simulations for the study of law systems, jurisdictional procedures, and judicial phenomena (3).

References

(1) Bianchini, D., Bono, C., Campi, A., Cappiello, C., Ceri, S., De Luzi, F., ... & Plebani, P. Challenges in AI-supported process analysis in the Italian judicial system: what after digitalization? Commentary paper. Digit. Gov.: Res. Pract. Just Accepted (October 2023). https://doi.org/10.1145/3630025.
(2) Faro S.: Scienze sociali computazionali, Diritto, Informatica giuridica; in Peruginelli G., Ragona M. (a cura di). L'Informatica giuridica in Italia: cinquant'anni di studi, esperienze e ricerche. Collana ITTIG, serie "studi e documenti" n.12, Edizioni Scientifiche Italiane, Roma (2014)
(3) Rezzani A.: Big Data: Architettura, tecnologie e metodi per l'utilizzo di grandi basi di dati, Apogeo Education, Roma (2013)

GEOREFERENCED SENTIMENT ANALYSIS OF TOURIST ATTRACTIONS IN THE CITY OF NAPLES

Luigi Celardo[1], Michelangelo Misuraca[2], Maria Spano[3]

[1] *University of Naples Federico II* (email:luigi.celardo@unina.it)
[2] *University of Calabria* (email: michelangelo.misuraca@unical.it)
[3] *University of Naples Federico* II (email:maria.spano@unina.it)

The growth of user-generated content (UGC) (1) in the travel industry makes reputation a significant part of travellers' daily activities. Digital platforms hosting reviews of attractions and other tourist points of interest, such as TripAdvisor, have become increasingly crucial for the management structure of the travel industry. Moreover, the reputation associated with tourism services is increasingly linked to material goods, atmosphere and the search for historical centres, which have become the main destinations of postmodern tourists. We propose calculating the polarity scores of reviews for all tourist attractions in Naples and using them in combination with other characteristics (e.g. duration of visit and type of site) to construct spatial clusters of tourist attractions. To explore narratives, we use the methodology termed G.R.A.S.S (Geo-Referenced Analysis of Sentiment Score)., which involves a process of web scraping reviews on the TripAdvisor site and then combining sentiment analysis and spatial clustering techniques such as agglomerative hierarchical clustering (AHC) with spatial constraint (3). Our results show geo-referenced attraction polarity scores with green and red areas according to positive or negative sentiment via a navigable interactive map (2). Furthermore, the possibility of using textual information as an additional variable for clustering also allows the identification of geographical areas that are more characterised by the type of attraction, the duration of the visit and the enjoyment of the experience to have a regionalisation of the city of Naples according to the travellers' experiences of the different attractions. The insights gained from this analysis are valuable for future research, especially for exploring and visualising tourism perceptions, providing a tool to monitor the attractiveness of an area and develop further actions or policies to improve the tourism sector.

References

(1) Baka, V., *The becoming of user-generated reviews: Looking at the past to understand the future of managing reputation in the travel sector*." Tourism management 53 (2016): 148-162.
(2) Celardo L., Misuraca M., Spano M. *Georeferencing sentiment scores to map and explore tourist points of interest*. In: 5th International Conference on Advanced Research Methods and Analytics (CARMA 2023). Editorial Universitat Politècnica de València; 2023.
(3) Murtagh, F., and Pedro C. *Algorithms for hierarchical clustering: an overview.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2.1 (2012): 86-97.

EXPLAINABLE (STATISTICAL) APPROACH TO NLP RECOMMENDER SYSTEM

Simone Travaglione[1], Roberta Siciliano[2]

[1] *University of Naples Federico II*
[2] *University of Naples Federico II* (email:roberta.siciliano@unina.it)

SUSTAINABILITY REPORTING IN ITALIAN CORPORATIONS: UNCOVERING GOVERNANCE DIVERSITY

Alessandra Belfiore[1], Corrado Cuccurullo[2]

[1] *University of Naples Federico II* (email: alessandra.belfiore@unina.it)
[2] *University of Naples Federico II* (email: corrado.cuccurrllo@unina.it)

Corporate financial reports have increasingly incorporated narrative sections, enriching their content and objectives. This development necessitates a shift in financial analysis from focusing solely on numerical data to integrating text analysis. Our study focuses on a comprehensive content analysis of sustainability reports from various Italian companies, examining the key topics and messages related to sustainability and Corporate Social Responsibility (CSR). To explore into corporate sustainability reporting, we employed a novel open-source content analysis tool named "Tall" (Text Analysis for all). This advanced tool specializes in both conceptual and thematic analysis. Moreover, its visual representation enables researchers to clearly understand and quantify the textual structure of the reports. Our findings reveal common themes related to corporate governance, including the roles of boards and auditors. Furthermore, we also discovered significant variations in how these themes are emphasized across different companies. This diversity suggests that the motivations and approaches to sustainability efforts vary considerably among firms. The insights gained from this analysis are valuable for future research, especially studies exploring the link between a company's sustainability practices and its overall organizational performance. By highlighting the thematic dimensions in sustainability reporting, our study offers a clearer understanding of how companies integrate and communicate sustainability in their financial reports, with a specific focus on the Italian corporate context.

PLS-SEM: A BIBLIOMETRIX TALE DEVELOPMENT

Enrico Ciavolino[1], Massimo Aria[2], Mario Angelelli[3]

[1] *University of Salento* (email: enrico.ciavolino@unisalento.it)
[2] *University of Naples Federico II* (email: massimo.aria@unina.it)
[3] *University of Salento* (email: mario.angelelli@unisalento.it)

This study provides a comprehensive analysis of the knowledge structure surrounding Structural Equation Modelling (SEM) based on the Partial Least Squares (PLS) estimator through systematic and reproducible bibliometric citation analysis. Utilizing the Bibliometrix package in R, the analysis encompasses 8,337 documents extracted from the Web of Science (WoS) database by Clarivate. The aim is to present a dynamic overview of PLS-SEM research activity, offering scholars an enriched understanding of its historical context, current status, and potential future directions. The findings reveal seminal papers, diffusion patterns across diverse research domains, and emerging applications. Moreover, the research addresses the increasing fragmentation of PLS-SEM-related fields, presenting results from key outcomes, including the identification of main research areas, the evolution of themes using trend topics, and the dynamic conceptual structure.

FORECAST VIRAL CONTENT ANALYSIS IN SOCIAL MEDIA

Domenica F. Iezzi[1], Roberto Monte[2], Daniele Pasquini[3]

[1] *Tor Vergata University of Rome* (email: stella.iezzi@uniroma2.it)
[2] *Tor Vergata University of Rome* (email: roberto.monte@uniroma2.it)
[3] *Tor Vergata University of Rome* (email: pasqualini@lettere.uniroma2.it)

Digital content's ability to become popular quickly and reach millions of people is called content virality. In recent years, there has been a growing number of studies on viral posts, given their ability to influence social, economic and political outcomes in contemporary society (1). Understanding the mechanisms that regulate virality can result fundamental in determining the popularity and involvement of central trends and topics in society (3).This paper aims to propose a model to predict whether a post will go viral or not. According to Kalra et al. (2), we can identify two virality approaches in the literature: 1) content-based popularity prediction (6; 5); and 2) circulation-based popularity prediction (4). In the first case, it estimates virality using textual and multimedia characteristics related to the content of a post. The second case models the network structure that connects users and studies how posts or news spreads among users. This research considers a mixed approach using sentiment analysis measures, key performance indices, and user popularity. We test our model using several social media (Instagram, Facebook, and TikTok) on different topics.

References

(1) Borges-Tiago M. T., Tiago, F., Cosme C., (2019). Exploring users' motivations to participate in viral communication on social media, *Journal of Business Rese*arch, Volume 101: 574-582.
(2) Kalra S., Kumar C. H. S., Sharma Y. and Chauhan G. S. (2022). Comparative Analysis of Various Machine Learning Based Techniques for Predicting the Virality of Tweets. In *12th International Conference on Cloud Computing, Data Science & Engineering* (Confluence), Noida, India, 601-605.
(3) Lopez Y.L., Grimaldi D., Garcia S., Ordoez J., Carrasco-Farre C., Aristizabal A.A. (2022). Artificial Intelligence Model to Predict the Virality of Press Articles. In Proceedings of the 2022 14th International Conference on Machine Learning and Computing (ICMLC '22). Association for Computing Machinery, New York, NY, USA, 221–228.
(4) Xu Z, Qian M. (2023) Predicting Popularity of Viral Content in Social Media through a Temporal-Spatial Cascade Convolutional Learning Framework. *Mathematics*. 11(14):3059.
(5) Rivas, L., Galea, M. (2019). Influence analysis for the generalized Waring regression model. *Journal of applied statistics*, *47*(1), 1–27.
(6) Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A.J., Olmo-Jiménez, M.J., & Martínez-Rodríguez, A.M. (2009). A generalized Waring regression model for count data. *Comput. Stat. Data Anal., 53*, 3717-3725.

MEASURING KNOWLEDGE DISTANCE: A SEMANTIC ANALYSIS OF SCIENTIFIC PUBLICATIONS

Luca D'Aniello[1], Nicolás Robinson-García[2], Massimo Aria[3], Corrado Cuccurullo[4]

[1] University of Naples Federico II (email:luca.daniello@unina.it)
[2] University of Granada (email: elrobin@ugr.es)
[3] University of Naples Federico II (email:massimo.aria@unina.it)
[4] University of Naples Federico II (email: corrado.cuccurullo@unina.it)

In recent years, the surge in textual data has led to a significant increase in information overload (1). The rapid pace of document publication within the scientific literature domain intensified this challenge. Given the crucial role of academic publishing in disseminating new knowledge and advancing scientific understanding, the publication of scientific documents allows sharing new discoveries and innovations among the scientific communities, contributing to further advancements and progress across different research fields (2). Since the evident growth of publications places scholars in a scenario inundated by the overwhelming volume of scientific material, this paper aims to understand if the expansion of scientific literature, in terms of the number of publications, is coherent with the growth of knowledge. Evaluating the growth of knowledge can be interpreted as the assessment of new information disseminated through the publication of scientific literature.

Measuring the extent of differentiation between publications can be accomplished through the use of a similarity distance technique, in order to analyze the degree of overlap and divergence between documents. To achieve this, we will perform the Word Mover's Distance (WMD), a semantic-based distance algorithm that quantifies the dissimilarity among texts of documents as the minimum distance that the embedded words of one document need to travel to reach the embedded words of another document (3).

The WMD analysis was performed on documents related to the h-index and its variations, including the r index, g-index, and a-index. These articles delve into several adaptations of these indexes, providing insights into both their strengths and weaknesses. After performing the text pre-processing phase and employing a word embeddings model, the algorithm generated a distance matrix, effectively quantifying semantic (dis)similarities within the texts. Consequently, a cluster analysis was employed on the similarity matrix to group similar documents and identify those with distinct characteristics. Our findings shed light on the intricate dynamics between the increase in scientific publications and the growth of knowledge. The cluster analysis based on the WMD similarity matrix provides a potential framework for assessing document distinctions and identifying different knowledge contributions within the scientific literature domain.

References

(1) Sarker A, Molla D, Paris C. Automated text summarisation and evidence-based medicine: A survey of two domains. arXiv preprint arXiv:1706.08162. 2017.
(2) Sánchez-García E, Martínez-Falcó J, Seva-Larrosa P, Marco-Lajara B. Delving into the analysis of scientific production and communication in academic literature. J Librarianship Inf Sci. 2024;0(0). https://doi.org/10.1177/09610006231223168.
(3) Kusner M, Sun Y, Kolkin N, Weinberger K. From word embeddings to document distances. In: International conference on machine learning. PMLR; 2015. p. 957-966.

INVESTIGATING THE IMMIGRATION DEBATE ON NEWSPAPERS: A STATISTICAL ANALYSIS OF MEDIA LANGUAGE

Alex Cucco[1], Emiliano del Gobbo[2], Lara Fontanella[3], Sara Fontanella[4], Annalina Sarra[5]

[1] *Imperial College of London* (email: a.cucco20@imperial.ac.uk)
[2] *University of Foggia* (email: emiliano.delgobbo@unifg.it)
[3] *University of Foggia* (email: lara.fontanella@unich.it)
[4] *Imperial College of London* (email: s.fontanella@imperial.ac.uk)
[5] *University of Chieti-Pescara* (email: annalina.sarra@unich.it)

In the contemporary societal landscape, immigration is a multifaceted and debated topic covering diverse ideological spectrums. The discourse on immigration, particularly within newspapers, reflects a divergence of varied opinions offered to the public. This study addresses the need to investigate and characterize the ongoing immigration debate, focusing on the contrasting perspectives prevalent in selected media outlets. Our primary aim is to examine the diversity in news coverage across various newspapers to understand the different perspectives presented to readers. Through the analysis of news articles from multiple sources, we aim to identify patterns, themes, and contrasting viewpoints that shape the narrative on various issues. This comprehensive exploration will offer insights into the range of opinions and biases inherent in media reporting. The proposed study employs statistical techniques, including network analysis of word co-occurrence, to investigate the underlying structures of the immigration discourse. This involves mapping relationships between words to identify patterns and connections, offering insights into prevailing opinions surrounding immigration. Network analysis applied to textual data in the context of news reporting can offer insights into the different positions and polarizations within a debate. By representing textual interactions as a network, nodes can be identified as words, and edges as connections or interactions between them. Network indices, such as centrality measures, can be employed to identify keywords or topics that play influential roles in shaping the debate. Differential network analysis can then be applied to compare semantic networks, revealing different shades of opinion related to a specific topic highlighting different ideological position. The examination of network structures can characterise the polarization of views among newspapers. In conclusion, the study emphasizes the central role of network analysis in decoding the complex space of opinions on immigration. By examining semantic networks, it provides insights that offer a comprehensive understanding of the diverse perspectives within newspapers.

CLUSTERING OF ATTRIBUTED NETWORKS VIA DISTATIS

Giancarlo Ragozini[1], Valeria Policastro[2], Roberto Rondinelli[3]

[1] *University of Naples Federico II* (email: giragoz@unina.it)
[2] *University of Naples Federico II* (email: valeria.policastro@unina.it)
[3] *University of Naples Federico II* (email: roberto.rondinelli@unina.it)

Community detection is one of the relevant tasks in the context of Social Networks and Complex Networks. Communities are cohesive subsets of nodes that are densely connected internally and sparsely connected externally, defined according to the topology structure of the network. Considering the different characteristics of the network (weighted/unweighted, directed/undirected, one-mode/two mode, etc.), several algorithms have been developed based on modularity, betweenness, random walk, etc. Although, these methods improved the resolution and accuracy of detected communities, the information around a network is even more complex and rich. Often, network datasets include additional information, such as attributes (features) of the nodes, which can be an efficient support to describe the communities and their configuration. Indeed, homophily, territorial proximity, counterfactual factors, spreading and conformism can affect the network formation process and therefore the composition of communities. In the last years, researchers posed the attention on this topic, trying to include the information of the nodes' attributes in the detection of the communities. For examples, one of these proposals modifies the hierarchical clustering algorithm including relational constraints; another approach describes the communities in the context of Subgroup Discovery; finally, another contribution provides a data-driven probabilistic method on multilayer networks. Suited also for single layer networks and for the prediction of missing links or attributes. In this framework, this work proposes the use of DISTATIS (a three-way multidimensional scaling) to combine simultaneously the information provided by the topological structure of the network and the attributes (quantitative and qualitative) of the nodes. To demonstrate the applicability and the good performance of this approach we simulated different type of networks, and different type of quantitative, and qualitative attributes which can be more and less related to the network structure. The first results detect a good compromise space of the three

A NEW APPROACH FOR ANALYSING FUNCTIONAL DEPENDENCIES NETWORK DATA

Elvira Romano[1], Andrea Diana[2], Antonio Irpino[3]

[1] *University of Campania "L. Vanvitelli"* (email: elvira.romano@unicampania.it)
[2] *University of Campania "L. Vanvitelli"* (email: andrea.diana@unicampania.it)
[3] *University of Campania "L. Vanvitelli"* (email: antonio.irpino@unicampania.it)

In today's interconnected world, network data play a crucial role in enabling communication and data transfer across various platforms and locations. Analysing such data poses challenges due to the several factors that must be considered. In the last years, several statistical models for network data have been proposed (1). The most common models includes: Graphical Models, Exponential Random Graph Models (ERGM), Spatial Models, Temporal Models. Each of them provides a different approach to analyse and interpret the underlying structure. Our work introduces two innovative concepts for data analysis in graph structured data, Network Functional Data (NFD) representing time series signals as functions on network nodes, and Network Weighted Functional Regression (NWFR) model exploring relationships between functional response variables and functional predictors in a weighted network. In addition a functional conformal approach is proposed to validate the defined model. Results of the proposed method are shown for a benchmark data set on real environmental data**.**

References

(1) Salter-Townshend, Michael, Arthur White, Isabella Gollini, and Thomas Brendan Murphy. "Review of Statistical Network Analysis: Models, Algorithms, and Software" 5, no. 4 (2012).
(2) Diana, A., Romano, E., Irpino A.: A new topological weighted functional regression model to analyse wireless sensor data. In: Proceedings of the Statistics and Data Science Conference. Pavia University Press, 2023, ISBN: 978-88-6952-170-6.

ANALYSIS OF MULTIMORBIDITY PATTERNS VIA GRAPHICAL MODELS

Erika Banzato[1], Giovanna Boccuzzo[2], Alberto Roverato[3]

[1] *University of Padova* (email: erika.banzato@unipd.it)
[2] *University of Padova* (email: Giovanna.boccuzzo@unipd.it)
[3] *University of Padova* (email: alberto.roverato@unipd.it)

The analysis of multimorbidity represents a crucial challenge for healthcare systems, given its close relation with adverse health outcomes, more complex clinical management, and associated costs. Multimorbidity, defined as the coexistence of two or more chronic diseases within an individual, necessitates a thorough investigation of disease associations to comprehend the underlying phenomenon. These findings not only help on generating new hypotheses on potential shared biological processes but also aid in quantifying the impact of multimorbidity on health-related outcomes and quality of life, thereby enhancing preventive measures. In this setting, graphical models play an important role in the identification of patterns and offer an intuitive tool to represent the multimorbidity network. Representing associations between variables as a graph, with each node denoting a disease and edges the relationships between them, provides an interpretable depiction of conditional dependencies. This study leverages a large dataset from a medical registry in the Padova province (Italy) to showcase how graphical models enhance our understanding of multimorbidity by revealing the connecting structure between variables. The initial phase involves identifying the associations' structure, followed by estimating the joint model associated with it, offering insights into the strength of these associations. Utilizing the estimated model as a foundation, we illustrate the possibility of clustering variables by identifying groups of highly interconnected sub-graphs. We also provide a sex-stratified analysis, to show how differences can influence both the structure and the strength of the associations. Finally, we integrate an analysis of outcomes measured in the following year, highlighting the incorporation of this approach into a comprehensive analytical framework.

CULTURES AS NETWORKS OF CULTURAL TRAITS: A UNIFYING FRAMEWORK FOR MEASURING CULTURE AND CULTURAL DISTANCES

Luca De Benedictis[1], Roberto Rondinelli[2], Veronica Vinciotti[3]

[1] *University of Macerata*
[2] *University of Naples Federico II* (email: roberto.rondinelli@unina.it)
[3] *University of Trento*

Using data from the sixth wave of the World Value Survey and operationalising a definition of national culture that emphasises both specific cultural traits and the inter dependence among them, this paper proposes a methodology to reveal the latent network structure of every national culture and to measure the cultural distance associated with every pair of countries as a Jeffreys' divergence between copula graphical models. The two components of this new measure of cultural distance show different correlations with measures of geographical, historical, economic, and political distance among countries and with the similarity in the topologies of the countries cultural networks.

DATA-DRIVEN MODEL BUILDING FOR LIFE-COURSE EPIDEMIOLOGY

Anne Helby Petersen[1], Claus Thorn Ekstrøm[2]

[1] *University of Copenhagen* (email: ahpe@sund.ku.dk)
[2] *University of Copenhagen* (email: ekstrom@sund.ku.dk)

2928 Danish men born in 1953 followed from birth until age 65 years with information from several contacts throughout their lives. Primary outcome is depression at early old age.
Life-course data are common in epidemiology where individuals are followed over time, and variables have a known partial temporal ordering. We propose a temporal extension of the PC algorithm for causal discovery of the underlying life-course model from observational data.
We develop the temporal PC algorithm (TPC) for incorporating temporal information in causal discovery with specific suggestions on how to use this when no oracle property is available, and statistical tests are needed to evaluate conditional independencies to learn the underlying causal structure. We propose a regression-based test as a necessary (not sufficient) criterion for conditional independence. Further, we show how a sequence of analyses with varying significance levels may be necessary to infer a "useful" life-course model with high retention rates for edges in the inferred graph.
Depression in early old age is found to be conditionally independent of all remaining variables given information about depression history in adulthood. Thus, there is no benefit in including childhood information, or other variables from adulthood, to understand why a depression develops in early old age. No causal effects of birth weight or birth length that span longer than youth are found. This is in contrast to myriad studies linking these factors to diabetes, death from ischemic heart disease, and mental health outcomes. Conclusions: We developed the temporal PC algorithm to produce life-course models from observed data. Information from the whole life course is considered jointly and allows for exploratory model building. This facilitates building global models that can provide empirical evidence about presence or absence of causal links between exposures occurring in different periods.

TEACHERS' BELIEFS ON THE USE OF DIGITAL TECHNOLOGIES AT SCHOOL: TEXT MINING OF OPEN-ENDED QUESTIONS

Annalina Sarra[1], Maila Pentucci[2], Eugenia Nissi[3]

[1] *University "G. d'Annunzio", Chieti-Pescara* (email: annalina.sarra@unich.it)
[2] *University "G. d'Annunzio", Chieti-Pescara* (email: maila.pentucci@unich.it)
[3] *University "G. d'Annunzio", Chieti-Pescara* (email: eugenia.nissi @unich.it)

In the era of digital plenitude (1), educational institutions face persistent challenges in aligning their strategies with technological advancements, adopting a learning-ecosystem perspective. Following the digital overload imposed by the pandemic, teachers consider technologies integration indispensable, but their extensive use in schools has also triggered attitudes of resistance (2). In this contribution, we aim to reflect on teachers' beliefs about the use, effectiveness and learning effects of digital, through the analysis of 898 responses to open-ended questions, collected through a questionnaire concerning digital learning ecosystems. To automate the analysis of the open-ended responses, we adopt a Structural Topic Model (STM; 3). STM, an extension of the LDA, integrates features from a correlated topic model and the sparse additive generative topic model. The underlying assumption of STM remains that documents stem from a combination of topics. However, the model permits correlations among topic proportions, enabling an exploration of how both the prevalence of topics and their associated content may vary based on covariates. The first results concern: A) the identification of some themes that clearly fit with what the recent literature highlights, including the interplay between digital education and early childhood, as well as the pressing issue of teacher training in technology. B) the possibility of clustering teachers' postures and attitudes according to their beliefs. The utilization text mining techniques on prominent themes in educational research has provided intriguing insights in the current context. On one hand, it has brought attention to the correlations between resistance to technological innovation and uncertainties regarding the new professional skills it demands. On the other hand, the exclusively data-driven approach has facilitated the illumination of certain latent aspects in teachers' perspectives, which may not have been as apparent in a question- and context-driven analysis.

References

(1) Bolter, J. D. (2019). *The digital plenitude: The decline of elite culture and the rise of new media*. Boston: MIT Press.
(2) Andriani, E., & Bram, B. (2022). Technology use in teaching literature amid and post pandemic: teachers' perceptions. *Premise: Journal of English Education*, 11(2), 329-347.
(3) Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013). The structural topic model and applied social science, in *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation,* Cambridge, MA.

THE SPREAD OF RANDOM FOREST ACROSS SCIENTIFIC RESEARCH FIELDS: A COMPREHENSIVE BIBLIOMETRIC ANALYSIS

Agostino Gnasso[1], Luca D'Aniello[2], Massimo Aria[3]

[1] *University of Naples Federico II* (email:agostino.gnasso@unina.it)
[2] *University of Naples Federico II* (email:luca.daniello@unina.it)
[3] *University of Naples Federico II* (email:massimo.aria@unina.it)

In recent years, the scientific production landscape has experienced a surge growth, evidenced by a notable increase in the volume of scientific documents published.
Currently, scientific documents are systematically indexed on bibliographic databases like Scopus and Web of Science. Among these, OpenAlex stands out as one of the most recently use bibliographic database, offering a free and open-source catalog of the global research system, with an impressive index exceeding 240 million documents.
Within the domain of computational statistics, the article on Random Forest, authored by Leo Breiman in 2001, counts over 80.000 citations. The Random Forest algorithm, an ensemble learning method that combines multiple decision trees to enhance prediction accuracy and mitigate overfitting. Bibliometric approaches play a crucial role in analyzing huge volumes of scientific document data through the application of mathematical and statistical methods.
This study aims to comprehensively analyze the whole collection of documents that cited the Random Forest article, by identifying the publication trends within different scientific research fields. Leveraging the OpenAlex classification system with Concepts, we aim to focus on specific research domains and applications where the Random Forest approach has been applied. Our objective is to delineate the trends and contexts in which Random Forest has been cited, providing a nuanced understanding of its widespread adoption. This exploration contributes valuable insights into the dynamic landscape of scientific inquiry and application across the scientific domains.

References

(1) Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.
(2) Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. ArXiv. https://arxiv.org/abs/2205.01833

ANALYZING OFFICIAL STATISTICS WITH SYMBOLIC DATA ANALYSIS

Paula Brito[1], Pedro Duarte Silva[2]

[1] *Universidade do Porto & LIAAD-INESC TEC* (email: mpbrito@fep.up.pt)
[2] *Católica Porto Business School & CEGE, Universidade Católica Portuguesa*

Symbolic Data Analysis (see, e.g., (1)) provides a framework for the representation and analysis of complex data, comprising inherent variability. That is the case when the entities under analysis are not single elements, but groups formed from the aggregation of the original statistical units. To this aim, new variable types have been introduced, whose realizations are not single real values or categories, but sets, intervals, or distributions over a given domain. This framework is of particular relevance in the analysis of official statistics, where the interest often lies in statistical units at a higher aggregated level, and where confidentiality issues prevent the dissemination and analysis of the microdata.

In this work, we consider data where individual units, resulting from the aggregation of large amounts of microdata, are described by distributions of numerical attributes. We propose parametric models for numerical distributional variables based on the representation of each distribution by a central statistic, and the logarithm transformation of inter-quantile ranges, for a chosen set of quantiles. Multivariate Normal distributions are assumed for the whole set of indicators, with alternative structures of the variance-covariance matrix. This model then allows for multivariate parametric analysis of distributional data. Applications to official data put in evidence the benefit of the proposed approach in the context of official statistics data analysis.

References

(1) Brito, P. Symbolic Data Analysis: Another Look at the Interaction of Data Mining and Statistics. WIREs

ANTICIPATING DELAYS IN COHESION INFRASTRUTURE PROJECTS. A MACHINE LEARNING APPROACH

Gianluca Monturano

University of Modena e Reggio Emilia (email gianluca.monturano@unimore.it)

Territorial fragilities in reaping the benefits of cohesion policies are partly due to difficulties in allocating and spending resources by central administrations. The efficiency of the allocative mechanism depends on planning and administrative capacities of beneficiary local authorities, project peculiarities, socio-economic dynamics, institutional characteristics, and the need for coordination among authorities. This study focuses on delays in implementing projects for territorial cohesion. The results reveal spatial concentration of delays in Italy's most fragile areas, influenced by historical issues affecting Italian growth.

NATURE-BASED SOLUTIONS AND PROXIMITY TOURISM: THE EXPERIENCE OF THE YOUNGER GENERATION

Luigi Bollani[1], Alessandro Bonadonna[2]

[1] *University of Torino* (email: luigi.bollani@unito.it)
[1] *University of Torino* (email: alessandro.bonadonna@unito.it)

The European Commission states the definition of Nature Based Solutions (NBS) as solutions "inspired and supported by nature, which are cost-effective, simultaneously provide environmental, social and economic benefits and help build resilience. Such solutions bring more, and more diverse, nature and natural features and processes into cities, landscapes, and in this context, this study aims to explore the relationship between NBSs and outreach tourism from the perspective of the younger generations. On the one hand, NBS may contribute to environmental, social and economic benefits and can increase proximity tourism in some areas; on the other hand, the younger generations are generally considered to be more sensitive to sustainability and sustainable actions and very interested in tourism and nature. The methodology used included a demoscopic survey to collect data and reached 988 university students living in Turin.
The results show that NBS initiatives are perceived by GenZ as important in terms of safeguarding and enhancing the cultural and natural heritage of the urban areas involved and can improve their tourism and recreational value.
This research is particularly relevant because it can help institutions consider a new approach to stimulate proximity tourism in and around the cities, enhancing NBSs as a possible attraction for GenZs.

Service quality matters: a new approach for assessing airline operational performance

Agnese Rapposelli[1], Stefano Za[2], Eusebio Scornavacca[3]

[1] *University of Chieti-Pescara, "G. d'Annunzio"* (email: agnese.rapposelli@unich.it)
[2] *University of Chieti-Pescara, "G. d'Annunzio"* (email: stefano.za@unich.it)
[3] *Arizona State University* (email: escornav@asurite.asu.edu)

In the service industry it is important to integrate the analysis of service productivity to aspects related to service quality. In this work, focused on 30 Italian airline domestic routes, efficiency measurement methods are adapted to airline industry by including an indicator of service quality (3) represented by the average delay.

The purpose of this study is to propose a novel efficiency evaluation approach that incorporates both technical efficiency and service quality, adopted for evaluating the operational performance of domestic routes in the airline industry.

The operational performance of domestic routes is empirically evaluated by employing a Principal Component Analysis - Data Envelopment Analysis (PCA-DEA) model, thus overcoming the traditional DEA approach (2), to identify the most efficient airline routes.

Incorporating PCA in DEA formulation allows to include several cost categories in routes' performance evaluation without losing information (1). The addition of this service indicator in our analysis improves the understanding of the efficiency results registered among routes. The results show that 9 out of 30 units analysed are on the efficient frontier and that several routes are operating at a high level of efficiency.

The proposed approach can assist decision-makers with the identification of sub efficient routes and pursue actions to improve levels of efficiency, such as potential closure of inefficient routes as well as opportunities for opening new ones.

References

(1) Agovino M and Rapposelli A (2013) Inclusion of disabled people in the Italian labour market: an efficiency analysis of law 68/1999 at regional level. *Quality & Quantity* 47(3): 1577–1588.

(2) Rapposelli A (2012) Route-based performance evaluation using Data Envelopment Analysis combined with Principal Component Analysis. In: Di Ciaccio A., Coli M. IJMA (ed.) *Advanced Statistical Methods for the Analysis of Large Data-Sets. Studies in Theoretical and Applied Statistics*. Berlin: Springer-Verlag, pp. 351–360.

(3) Suzuki Y (2000) The relationship between on-time performance and airline market share: a new approach. *Transportation Research Part E* 36(2). Elsevier Ltd: 139–154.

CUSTOMER SATISFACTION IN RAIL TRANSPORT

Pietro Iaquinta[1], Eveny Ciurleo[2]

[1] *University of Calabria* (email: pietro.iaquinta@unical.it)
[2] *University of Calabria* (email: eveny.ciurleo@unical.it)

Customer satisfaction in the rail transport sector is a critical element that can impact their subjective well-being, general quality of life (Erik Bjørnson Lunke, 2020) as well as the sustainability of rail networks.
In this work we will try to investigate the factors that influence customer satisfaction in the context of regional rail transport in Italy, which has been the subject of numerous studies over time; however, there is a growing need for integrated approaches that consider sociodemographic variables, usage patterns and infrastructure aspects to gain a more comprehensive and detailed understanding; to do this we will try to answer the following question: "To what extent do variables such as the transport use index, the gender indicator and the type of railway network influence the degree of customer satisfaction in regional rail transport?"
We will use a linear regression approach to explore the relationships between the degree of satisfaction (independent variable) and the dependent variables, including the transport use index, the gender indicator and the types of rail network.
We expect that the analysis will reveal significant differences in satisfaction with the variables used in a standard manner.
The contribution of this research could be important for railway operators and local authorities in order to optimize customer satisfaction and improve the quality of the service offered. Furthermore, it is expected that this investigation will help fill the current gaps in the integrated understanding of the key factors influencing customer satisfaction in the rail transportation industry. It is also intended to make a significant contribution to the existing literature by offering a comprehensive perspective on customer satisfaction in the context of regional rail transport.

PROCESS MINING DISCOVERY STARTING FROM PROCESS-UNAWARE DATA: LESSONS LEARNED FROM AN APPLICATION IN HEALTHCARE

Simone Leonetti[1], Andrea Burattin[2], Domenico Tricò[3], Nicolai Skytte Mikkelsen[4], Qixin Ma[5], Fuad Hassan Jama[6], Paulo Emanuel Ferreira Lima[7], Chiara Seghieri[8]

[1] *Sant'Anna School of advanced studies* (email:s.leonetti@santapisa.it)
[2] *DTU Technical University of Denmark* (email:andbur@dtu.dk)
[3] *University of Pisa* (email:domenico.trico@unipi.it)
[4] *DTU Technical University of Denmark*
[5] *DTU Technical University of Denmark*
[6] *DTU Technical University of Denmark*
[7] *DTU Technical University of Denmark*
[8] *Sant'Anna School of advanced studies* (email: c.seghieri@santannapisa.it)

Healthcare processes are complex, highly flexible, and multidisciplinary. Due to their individuality and specificity, the unstructuredness of these processes is reflected in the high variability of the logs data recorded when they are executed.
Process mining (PM) techniques (1), control-flow discovery in particular, aim to extract valuable insights from event logs – describing the sequence of activities that were performed – to automatically construct process models. The existing PM literature predominantly focuses on medical treatment processes andorganizational processes (2). Medical treatment processes involve diagnosis, therapies, and treatments delivered to patients, while organizational processes focus on knowledge coordination among professionals, unit management, and patient scheduling. Care providers are increasingly turning to systematic analysis of healthcare data to (re-)design processes, aiming to streamline care delivery, reduce costs, and enhance quality. Despite several process mining approaches that have been proposed in the healthcare domain, conducting process mining on process-unaware healthcare data remains an open challenge.
The objective is to highlight the challenges associated with the application of PM analysis in the context of healthcare non-aware process data.
We used MIMIC-IV dataset (3) as example of PM application to real-life healthcare data due to its public availability and its structure, which incorporates health data from various hospital sources, including emergency department data, patients' vital signs, and inter-departmental transfers. To address specific care tailored to each disease, we focused on a cohort of patients with myocardial infarction as the primary hospital diagnosis. The complexity of the individual patients' journeys is abstracted at an exploratory level, using PM methods, and focusing on hospital admissions, transfers, and the discharge processes.
The results reveal a high complexity of the patient journey in the hospital, reflecting the considerable variability present in healthcare non-aware process data. Further research is necessary to enhance the comprehensibility of the process model in the healthcare domain.

References

(1) W. van der Aalst et al., 'Process Mining Manifesto', 2012, pp. 169–194.
(2) E. Rojas, J. Munoz-Gama, M. Sepúlveda, and D. Capurro, 'Process mining in healthcare: A literature review', J. Biomed. Inform., vol. 61, pp. 224–236, Jun. 2016, doi: 10.1016/j.jbi.2016.04.007.
(3) A. E. W. Johnson et al., 'MIMIC-IV, a freely accessible electronic health record dataset', Sci. Data, vol. 10, no. 1, p. 1, Jan. 2023, doi: 10.1038/s41597-022-01899-x.

Exploring HIV hotspots in Zambia and Zimbabwe: an analysis using the INLA-SPDE approach

Micaela Arcaio[1], Anna Maria Parroco[2], Chibuzor Christopher Nnanatu[3]

[1] *University of Palermo* (email:micaela.arcaio@unipa.it)
[2] *University of Palermo* (email:annamaria.parroco@unipa.it)
[3] *University of Southampton*

In 2021, 25.6 million people living with HIV/AIDS resided in Sub-Saharan Africa (1). HIV is still a relevant topic, which explains why 10 of the 17 SDGs include actions and targets which can aid the most affected populations. The literature shows that geographical factors facilitate transmission, and living in high prevalence communities significantly increases the probability of individuals to become HIV positive as well (2). The aim of this preliminary study is to investigate geographical patterns of the prevalence of HIV in Zambia and Zimbabwe, using data from the Demographic and Health Survey. Indeed, this survey provides both the HIV status of respondents – determined by ELISA-type tests – and the geolocalisation of its clusters (of households). The data were used in a Bayesian hierarchical model within the framework of INLA-SPDE (3). These models allow to model HIV prevalence at the cluster level – which is assumed to be continuous in an underlying Gaussian field and with a Beta distribution. One model was estimated for each country, and post-estimation results were used to investigate geographical patterns of the prevalence of HIV. These values are then represented in the maps, which show several hotspots along the borders particularly for Zambia and Zimbabwe. Relevant hotspots are found in both countries, especially along the borders.
The results may highlight the effect of seasonal migration across borders on HIV diffusion. Moreover, using these methods could allow policy-makers to create programs that are measured to the people living in those areas, thus taking ethnicities, religious backgrounds, and cultural practices of their residents into consideration.

References

(1) NAIDS. Aidsinfo.com. [Cited 2024 Jan 15], from aidsinfo.unaids.org
(2) Messina JP, Emch M, Muwonga J, Mwandagalirwa K, Edidi SB, Mama N, Okenge A, Meshnick SR. Spatial and socio-behavioral patterns of HIV prevalence in the Democratic Republic of Congo. Social science & medicine. 2010 Oct 1;71(8):1428-35.
(3) Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2009 Apr;71(2):319-92.

COMPARISONS OF DIFFERENT METHODS TO DERIVE MRI-BASED AGING CLOCK FOR THE HEART AND EVALUATION OF THE EFFECT OF SOCIOECONOMIC INEQUALITIES ON THE AGE GAP

Lorenzoni Valentina[1], Andreozzi Gianni[2], Pier Giorgio Masci[3]

[1] *Sant'Anna School of advanced studies* (email: valentina.lorenzoni@santannapisa.it)
[2] *Sant'Anna School of advanced studies* (email: gianni.andreozzi1@santannapisa.it)
[3] *Kings College London* (email: pier_giorgio.masci@kcl.ac.uk)

Biological age (BA) is the term used to name an estimation of an individual's actual age, made up of age related biomarkers into a synthetic measure. BA and aging clocks (AC) have the potential to capture the actual aging process of the body and thus early detect process that may lead to the development of age-related diseases, such as cardiovascular and metabolic diseases (1,2). Advanced statistical modelling may serve to properly estimate those measures (3); moreover, this branch of research represents an intriguing field that may serve to understand socioeconomic inequalities in health (4).
The aim of this work is to evaluate different models to derive MRI-based AC for the heart and to assess the effect of socioeconomic condition (SEP).
Considering more than 39,000 subjects enrolled in the UK-Biobank (5) project for which MRI features were available, the Klemera-Doubal (KD) formula (6), conventional statistical methods (i.e. PCA) and machine learning approaches (i.e. LASSO regression), were used to develop MRI-based AC. Performance of methods was evaluated using the mean absolute error (MAE), plausibility of AC was assessed considering the range of values obtained, the degree of deviation and the correlation with chronological age using the Pearson correlation coefficient. Finally differences according to SEP assessed using independent sample T-test.
All methods resulted in AC significantly and highly correlated with chronological age (correlation values being between 0.6 and 0.8) and the MAE was lowest for KD and LASSO regression (3.5 and 4.9 respectively). For all AC and for both sexes the age gap (difference between AC and chronological age) resulted statistically significantly high among low educated subjects (p-value<0.001 for all).
The use of proper statistical methods may drive the road for research in the field of BA allowing also disentangle the role of aging and SEP.

References

(1) Ferrucci L, Gonzalez-Freire M, Fabbri E, Simonsick E, Tanaka T, Moore Z, Salimi S, Sierra F, de Cabo R. Measuring biological aging in humans: a quest. Aging Cell. 2020;19: e13080.
(2) Ferrucci L, Levine ME, Kuo PL, Simonsick EM. Time and the Metrics of Aging. Circ Res. 2018;123:740-744.
(3) Rutledge J, Oh H, Wyss-Coray T. Measuring biological age using omics data. Nat Rev Genet 2022; 23(12):715-727.
(4) Stringhini S, Carmeli C, Jokela M, Avendano M, Muenning P, Guida F et al. Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: a multicohort study and meta-analysis of 1·7 million men and women. The Lancet 2017; 389 (10075):1229-1237.
(5) Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015;12(3):e1001779
(6) Klemera P, Doubal S. A new approach to the concept and computation of biological age. Mech Aging Dev 2006; 127(3):240-8.

DATA-DRIVEN DECISIONS: OPTIMIZING EMERGENCY ROOM OPERATIONS FOR BETTER

Simone Paesano[1], Dario Sacco[2], Gaetano Gubitosa[3], Angela Annecchiarico[4], Federica D'Agostino[5], Valentina Di Palma[6], Maria Gabriella Grassia[7], Giuseppe Signoriello[8]

[1] *University of Naples Federico II* (email:simone.paesano@unina.it)
[2] *University of Naples Federico II* (email:Dario.sacco@unina.it)
[3] *A.O.R.N. Caserta*
[4] *A.O.R.N. Caserta*
[5] *A.O.R.N. Caserta*
[6] *A.O.R.N. Caserta*
[7] *University of Naples Federico II* (email: mariagabriella.grassia@unina.it))
[8] *University of Campania Luigi Vanvitelli* (email:Giuseppe.signoriello@unicampania.it)

In emergency room (ER) settings, the intricate relationship between patient waits times and the accuracy of triage codes is pivotal for ensuring effective patient care and outcomes. Wait times denote the interval from when patients arrive at the ER to when they receive medical evaluation or treatment, whereas triage codes categorize patients based on the urgency of their conditions, guiding priority in treatment. The dynamic between wait times and triage accuracy critically affects both the quality of care and patient satisfaction levels. Research underscores a significant link between the duration of wait times and the precision of triage assessments in ERs. It has been found that roughly 50% of triage evaluations in emergency settings may be inaccurate (1). Properly identifying patients with high-urgency needs is crucial for safeguarding patient safety, while accurate recognition of low-urgency cases enhances ER throughput and reduces wait times (2). Additionally, evidence points to a direct correlation between shorter wait times and improved patient outcomes, highlighting the critical nature of prompt medical attention (3). The goal of this study is to propose a new approach for analyzing patient waiting times, with a particular focus on deciphering the impact of inaccurate code allocations. By examining the textual congruence among diagnoses associated with variably colored codes, we aim to reveal any discrepancies between the assigned diagnoses and the corresponding triage codes. To achieve this, the contribution utilizes data related to the patient flow from the Emergency Department of the Sant'Anna and San Sebastiano Hospital (Caserta). Through this approach, the research seeks not only to identify mismatches between diagnostic assessments and triage categorizations but also to propose solutions to enhance the efficiency and accuracy of emergency medical services.

References

(1) Çetin, S., Eray, O., Cebeci, F., Coşkun, M., & Gözkaya, M. (2020). Factors affecting the accuracy of nurse triage in tertiary care emergency departments. Turkish Journal of Emergency Medicine, 20(4), 163. https://doi.org/10.4103/2452-2473.297462.

(2) Göktüğ, A., İlarslan, N., Vatansever, G., Özdemir, İ., Polat, O., Oğuz, A., … & Tekin, D. (2022). Evaluation of the validity and reliability of ankutriage, a new decision support system in pediatric emergency triage. Pediatric Emergency Care, 39(1), 28-32. https://doi.org/10.1097/pec.0000000000002750.

(3) Nikzadian, M., Wagner, J., Beiranvand, R., & Khormehr, M. (2022). Using weibull model of survival analysis workflow and its relevant factors: a prospective cohort study. Journal of Emergency Practice and Trauma, 9(1), 44-51. https://doi.org/10.34172/jept.2022.34.

PRIN The future of sustainability-P2022B3NFH
Contributo finanziato con fondi Missione 4 - Istruzione e Ricerca – Prin PNRR

MODELLING THE TOPICS OF ITALIAN TWEETS ABOUT THE 2022 ENERGY CRISIS

Matteo Farnè[1], Laura Zavarise[2]

[1] *University of Bologna* (email: matteo.farne@unibo.it)
[2] *University of Bologna*

During 2022, the energy crisis has emerged as a predominant subject in public discourse, extensively covered across various media outlets. Social networks have played a pivotal role in this context, providing individuals with a platform to actively engage in and contribute to the ongoing debate. This study examines a collection of Italian language messages, shared on Twitter, to comprehend platform users' perceptions of the energy crisis and its related aspects. Our focus is primarily on identifying the most debated topics, achieved through the application of unsupervised topic modelling. A total of five salient topics are identified and then categorized into themes, including energy markets dynamics, Ukraine crisis, energy supplies, Italian politics and household energy consumption.

FRANCYBAS: SMART STICK FOR FOREST SAFETY, TRAIL DISCOVERY AND BIODIVERSITY PROTECTION

Marco Colucci[1], Sonia Romani[2], Girolamo Radosti[3], Vito Santarcangelo[4]

[1] *L'Antincendio Srl* (email:amministrazione@lantincendio.it)
[2] *iInformatica Srl*
[3] *iInformatica Srl*
[4] *iInformatica Srl*

This research paper presents the patented Francybas smart stick for forest safety featuring an auger to conduct undergrowth sampling for fire prevention. Francybas thanks to the use of appropriate environmental and tracking sensors favores the recording of routes taken by users and the related discovery of new trails through data analysis techniques. The same principles, thanks to green recharging systems can be used to carry out dispersed detection. Furthermore, thanks to its drilling properties, the same stick can be used to dispense territorial seeds, preserving and promoting the protection of biodiversity.

References

(1) Zweifel R., Babst F. et al. (2023). Networking the forest infrastructure towards near real-time monitoring. *Science of The Total Environment*
(2) Giacalone, M. , Calciano M.V., Santarcangelo V., et al. (2021). A Novel Big Data Approach for Record and Represent Compliance in the Covid-19 Era. *Big Data Research*

ALLINCHAIN : CREATION AND VERACITY ANALYSIS OF A DISTRIBUTED BLOCKCHAIN IN SMART BINS FOR CERTIFIED DESTRUCTION

Giuseppe Stella[1], Giuseppe Oddo[2], Michele Di Lecce[3], Matteo Trimarchi[4], Diego Carmine Sinitò[5]

[1] *Stella All in One Srl* (email: giuseppe@dittastella.it)
[2] *iInformatica Srl*
[3] *iInformatica Srl*
[4] *iInformatica Srl*
[5] *iInformatica Srl*

This research paper recounts the design and implementation of a new blockchain called "All in Chain" applied to certified document destruction bins (patented) called Tekbin. This revolutionary "permissioned-consortium" approach sees through the Go-ethereum (aka Geth) client the implementation of a real model that guarantees integrity and transparency about activities provided by Tekbin smart bins. Thus, a revolutionary approach that sees the potential of Rpi in synergy with Arduino and the simplicity of permissioned distributed blockchain being exploited following the perspective of sustainable software engineering that sees optimizing energy consumption and exploiting the potential of the permissioned-consortium approach as a real revolution to be applied in multiple cases. Through the analysis of the collected data (OLTP) and the number of ledger bins, it is also possible to set up a mathematical evaluation of the blockchain's assurance performance, thus avoiding evasive cases and providing for external feedback to the user regarding the veracity performance of the data in it.

References

(1) Donvito, V., Schiavone, O., Santarcangelo V., Massa E. (2021). Big data for corporate social responsibility: blockchain use in Gioia del Colle , *Quality & Quantity*; **55(6):**1945-1971 Giacalone, M. , Calciano M.V., Santarcangelo V., et al. (2021). A Novel Big Data Approach

VARIABLE IMPORTANCE IN RANDOM FORESTS WITH GLOBAL SENSITIVITY ANALYSIS: A NUMERICAL EXPERIMENT

Giulia Vannucci[1], Roberta Siciliano[2], Andrea Saltelli[3]

[1] *University of Naples Federico II*
[2] *University of Naples Federico II* (email:roberta.siciliano@unina.it)
[3] *University Pompeu Fabra – Barcelon; University of Bergen*

In statistical models analysts are interested in both predicting the response variable conditionally on a set of covariates, and in understanding which covariate contributes the most to the variation of the response (1). In the Machine Learning (ML) literature, a clear distinction about the predictive and the explanatory purposes is generally made (2). Still an active area of investigation is to test whether a ML model, that by construction pursues predictive purposes, can also give some information about the data generating process. Achieving this extra bit of information would improve the interpretability of ML algorithms, that is crucial when such models are used, for example, for variable selection (3). The purpose of this work is to apply global sensitivity analysis (GSA) (4) to ML to explore (i) whether we confirm findings of standard ML feature selection algorithms, (ii) whether any algorithmic improvement is obtained (e.g. in term of speed of computation), and finally (iii) to explore whether we improve the interpretability of the results. We propose an algorithm based on GSA approach to rank regressors in terms of their importance in a random forest model (5). The approach makes use of a global sensitivity measure to act on binary variables (triggers) that activate the presence / absence of features. Then, an estimator of the total sensitivity index (6) is computed to obtain the new variable importance. Numerical experiments on simulated data shows that for some particular data generating process our approach rank the explanatory variables according to the data generating process itself. A comparison with other ML methods is explored in terms of difference of ranks regressor. The proposed approach is promising. When the interest of the researcher is to employ a ML method to make conclusions in terms of data generating process, the GSA instrument applied to ML methods can help in interpretability of results.

References

(1) Gottard, A., Vannucci, G., and Marchetti, G. M. (2020). A note on the interpretation of tree- based regression models. Biometrical Journal, 62(6), 1564-1573.
(2) Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science 16(3), 199–231.
(3) Rothacher, Y., and Strobl, C. (2023). Identifying informative predictor variables with random forests. Journal of Educational and Behavioral Statistics, 10769986231193327.
(4) Saltelli, A. (2002). Sensitivity analysis for importance assessment. Risk analysis, 22(3), 579- 590.
(5) Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.
(6) Saltelli, A., M. Ratto, T. H. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola (2008). Global sensitivity analysis: the primer. John Wiley.

EXPLORING SYNERGY: GENERATIVE ARTIFICIAL INTELLIGENCE FOR COMMUNICATING IN-DEPTH PROFILING OF PSYCHOGRAPHIC CLUSTERS CREATED WITH THE "THÉMASCOPE" APPROACH

Furio Camillo

University of Bologna (email: furio.camillo@unibo.it)

This presentation explores the innovative use of generative artificial intelligence in providing detailed cluster descriptions derived from psychographic cluster analysis. By applying advanced artificial intelligence models, we will illustrate how generative artificial intelligence can improve the understanding of the distinctive features within individual clusters, offering an in-depth and nuanced perspective especially useful for communicating clustering results. Case studies derived from applications to micro-marketing, CRM and hyper-profiled communication problems will be presented, in order to highlight the effectiveness of a specific way of generating prompts for requests to AI tools. The work will tend to highlight the potential for integration between psychographic analysis and emerging technologies for a more complete and detailed vision of diversity within natural groups defined and studied starting from psychographic investigations. The reference statistical methodology is that suggested by L. Lebart in 1989 (4), in reference to the "thémascope" approach typical of the French strategy of data analysis applied to large surveys.

References

(1) Camillo F. “L’immaginario psicografico dei giovani italiani nell’utilizzo dei social network” , MK ABI, n.6 2023
(2) Gu, Jindong, et al. "A systematic survey of prompt engineering on vision-language foundation models." arXiv preprint arXiv:2307.12980 (2023).
(3) Liu, Vivian, and Lydia B. Chilton. "Design guidelines for prompt engineering text-to-image generative models." *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022.
(4) Lebart, L.: Stratégies du traitement des donnees d’enquetes. La Revue de MODULAD, 3, 21–29, 1989

ADVANCED RADIOMICS AND MACHINE LEARNING IN ORAL CANCER DIAGNOSIS

Pasquale Piombino[1], Emanuele Carraturo[2], Cristiana Germano[3]

[1] *A.O.R.N. Caserta* (email: piombino@unina.it)
[2] *University of Naples Federico II* (email: emanuele.c2971995@gmail.com)
[3] *University of Naples Federico II*

Radiomics in oral cancer, particularly in the case of oral tongue squamous cell carcinoma (OTSCC), involves advanced imaging techniques combined with machine learning models to extract a vast number of features from MRI images. These features can range from the shape and size of the tumor to more complex textural and wavelet features. The process begins with image acquisition, typically using T1-weighted and T2-weighted MRI, followed by meticulous segmentation of the tumor area by expert radiologists. Preprocessing steps are applied to ensure uniformity across images, such as denoising, correction of intensity non-uniformities, intensity standardization, and voxel size resampling.

ONE DIGITAL HEALTH: ONE WORLD, ONE VISION, ONE ECOSYSTEM

Oscar Tamburis

National Research Council (email: oscar.tamburis@unina.it)

One Digital Health (ODH) proposes a unified framework for future health ecosystems (1). The current approach to health, which separates human health, animal health, and the management of the surrounding environment, is outdated and ineffective. ODH introduces a novel approach that takes into account the interconnectedness of these three areas.

The One Digital Health framework is built around two keys (One Health and Digital health), three perspectives (individual health and well-being, population and society, and ecosystem), and five dimensions (citizens' engagement, education, environment, human and veterinary healthcare, and Healthcare Industry 4.0). It aims to transform future health ecosystems by leveraging digital technologies, which can help us collect and analyze data more effectively, thus leading to better health outcomes – for example, digital technologies can help us track disease outbreaks in real-time, which can help us respond more quickly and effectively.

The overarching goal of One Digital Health is to digitally revolutionize future health ecosystems by implementing a systemic health and life sciences approach. This approach is meant to allow future generations of health informaticians to navigate the intrinsic complexity of novel health and care scenarios within digitally transformed health ecosystems. In this evolving landscape, citizens and their health data assume a pivotal role in managing individual and population-level perspective data. The primary challenges involve fostering efficient interactions and delivering near-real-time, data-driven contributions in systems medicine and systems ecology by bridging One Health and digital health communities. Digital health literacy is imperative, encompassing the capacity for understanding and engaging in health prevention activities, self-management, and collaboration in systemic, ecosystem-driven public health and data science research. Within a robust One Digital Health ecosystem, individuals must adopt an active and robust approach via the implementation of FAIR-compliant ODH Interventions, to prevent and manage health crises and disasters (2,3).

References

(1) Benis A, Tamburis O, Chronaki C, Moen A. One digital health: a unified framework for future health ecosystems. Journal of Medical Internet Research. 2021 Feb 5;23(2):e22189.

(2) Tamburis O, Benis A. One digital health for more FAIRness. Methods of Information in Medicine. 2022 Dec 3;61:e116-24.

(3) Benis A, Haghi M, Deserno TM, Tamburis O. One Digital Health Intervention for Monitoring Human and Animal Welfare in Smart Cities: Viewpoint and Use Case. JMIR Medical Informatics. 2023 May 19;11:e43871.

IN THE EYES OF EXPERTS: CLINICIANS' ASSESSMENT OF AI'S ROLE IN HEALTHCARE

Dario Sacco[1], Maria Gabriella Grassia[2], Liliana Massa[3], Salvatore Massa[4], Francesca Paola Pastena[5], Simone Paesano[6]

[1] *University of Naples, Federico II* (email:dario.sacco@unina.it)

[2] *University of Naples, Federico II* (email:mariagabriella.grassia@unina.it)

[3] *Gruppo San Donato (*email:liliana.massa@grupposandonato.it)

[4] *A.O.R.N. Caserta (*email:daysurgery@ospedale.caserta.it)

[5] *Campus Bio-Medico of Rome (*email:fp.pastena@alcampus.it)

[6] *University of Naples, Federico II* (email:simone.paesano@unina.it)

The purpose of the present study is to investigate the opinion of medical and healthcare professionals regarding the implementation of artificial intelligence (AI) in the hospital, surgical, and sociomedical settings. This was achieved by studying the causal relationships hypothesized within the family of validated models known as the Unified Theory of Acceptance and Use of Technology. These models are designed to explore the degree of use and acceptance of a new technology when it is introduced in a work setting. To achieve this goal, a survey was conducted targeting a specific sample, the medical personnel associated with S.I.C.A.D.S. (Italian Society of Ambulatory Surgery and Day Surgery). Through the administration of a survey, latent constructs believed to be determinants of AI use in healthcare were studied, and through the use of the Partial Least Squares - Structural Equation Modeling (PLS-SEM) model, these constructs were measured, and theoretically hypothesized causal relationships were estimated. Regarding the opinions of healthcare professionals, openended questions were used to collect qualitative data. Thus, the research objective was achieved by following a two-step strategy: first, by estimating the causal links underlying the factors that determined the acceptance of AI, followed by analyzing the textual content of the questionnaire through text-mining techniques. In particular, the Polarity Detection technique was employed to determine whether the opinion on AI was positive or negative. Next, the sample was divided according to the polarity of opinion on AI, and a multi-group analysis (MGA) was performed using the PLS-SEM-MGA model to investigate any significant differences in model coefficients among users with differently polarized opinions. The research identified significant considerations, particularly in contexts where AI is considered to be particularly effective, such as the diagnostic pathway, the clinical and surgical pathway, and patient monitoring and follow-up.

References

(1) Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., ... & Albekairy, A. M. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education*, *23*(1), 689.

(2) Haleem, A., Javaid, M., & Khan, I. H. (2019). Current status and applications of Artificial Intelligence (AI) in medical field: An overview. *Current Medicine Research and Practice*, *9*(6), 231-237.

(3) Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly*, 425-478.

PRIN The future of sustainability-P2022B3NFH
Contributo finanziato con fondi Missione 4 - Istruzione e Ricerca – Prin PNRR

THE PLATFORMIZATION OF CONSUMER CULTURE ON TIKTOK: THE #SHOECHALLENGE CASE

Alessandro Caliandro[1], Lucia Bainotti[2]

[1] *University of Pavia* (email: alessandro.caliandro@unipv.it)
[2] *University of Amsterdam* (email: l.bainotti@uva.nl)

With this contribution we look at how TikTok's templates and algorithmic logics are incorporated into users' everyday practices of content production as well as of interaction within the platform itself. In doing so, we analyse how TikTok's architecture prompts practices of ephemeral consumption, here intended as forms of ephemeral digital consumption (rather than other forms of fast-paced and temporary consumption, such as fast fashion). By mixing hashtag analysis, sound analysis, and the visual analysis of TikTok videos, this contribution illustrates how platform affordances can stimulate the emergence of ephemeral consumption practices. By focusing on one TikTok challenge, the #shoechallenge, results show that ephemeral consumption on TikTok is characterised by: a) the ubiquitous display of consumption; b) the limited temporality of video clips; c) the situational nature of users' performances; and d) the attempts at attention-seeking in an algorithmically mediated and memetic platform.

TIKTOK AND THE CHINESE DIGITAL FORM

Adam Arvidsson[1], Camilla Volpe[2]

[1] *University of Naples Federico II* (email: adamerik.arvidsson@unina.it)
[2] *University of Naples Federico II* (email: camilla.volpe@unina.it)

This paper looks at the 'creator economy' to develop the concept of a Chinese digital form. We define a 'digital form' as a set of affordances and practices that have consolidated around digital technologies as a result of nationally specific trajectories of development. The Silicon Valley digital form was built on an alliance between financial capital and consumer interests and centred on the practice of branding or 'Instafame'. Tiktok instead derives from a Chinese trajectory of platform development. It is part of a digital form that promotes more industrious creator practices and that replaces the branding logic of social media publics with a 'despotic' management of tastes and preferences through algorithms and AI. Such industrious and despotic qualities have deep roots that can be traced to the dynamics of Chinese market transition as well as possibly further back to reflect aspects of classical Chinese political economy.

THE MACHINE HABITUS AS METHOD. RESEARCHING CONTENT WITHOUT CONTEXT ON TIKTOK

Vincenzo Luise

University of Naples Federico II (email:vincenzo.luise@unina.it)

TikTok, a rapidly growing social media platform, is popular for its short videos, especially among the younger generation. It's often referred to as a 'meme machine' due to its emphasis on mimetic repetition over individual creativity. The platform's algorithm is known for its remarkable adaptability to user preferences. Unlike Instagram, the content users see on TikTok is primarily based on their interactions with the app, not their social network or hashtags. This creates an experience of interacting with an algorithm that provides 'content without context', rather than a community of users with shared interests. Researchers can collect and analyze TikTok data using the official research API, browser extensions, and web- scraper tools. However, these tools struggle to map the unpredictable and ever-changing TikTok user base without social anchors like follower networks or hashtags, which are common on platforms like Instagram. To overcome this challenge, we developed a methodological strategy within the digital ethnography approach. Our focus is on studying the digital micro-entrepreneurs in Naples. We call this strategy the "Machine Habitus as method", inspired by Airoldi's study on the algorithmic mechanisms that link humans with artificial social agents. By continuously interacting with the platform and providing it with specific feedback, we can guide the algorithm to present us with content that is relevant to our research. This method not only enhances our understanding of this specific demographic but also provides a blueprint for studying other social phenomena on TikTok.

DYNAMICS OF VIRAL TRENDS: A COMPREHENSIVE ANALYSIS OF SECOND-HAND AND VINTAGE CONTENT ON TIKTOK

Rocco Mazza[1], Marino Marina[2], Camilla Volpe[3], Davide Torre[4]

[1] *University of Bari Aldo Moro* (email:rocco.mazza@uniba.it)
[2] *University of Naples Federico II* (email:marina.marino@unina.it)
[3] *University of Naples Federico II* (email:camilla.volpe@unina.it)
[4] *University of Naples Federico II* (email:davide.torre@unina.it)

Viral phenomena are a crucial subject of study across various professional domains, spanning both humanities and sciences, and serving diverse purposes. This work aims to map the trajectory of a viral trend on TikTok, specifically focusing on second-hand and vintage content. These contents prove to be contemporary and essential, particularly from a socioeconomic perspective. Our study takes into consideration the public of users particularly interested in vintage and second-hand clothing, which deploy these trends to showcase their uniqueness and environmental consciousness. The platform examined represents the most recent generation of social networks, where an audience with new social dynamics and affordances is emerging. The well-known field in which we operate is based on the industrious production of content and social innovation. These dimensions emerge from the analysis of the Key Performance Indicators (KPIs), which highlight the importance of specific parameters for a particular video or video group. The methods adopted is the dynamic factorial analysis, this incorporates multiple procedures aimed at observing the temporal development of the viral phenomenon. The technique involves, firstly, a principal component analysis with a synthesis measure to represent the case plane, alongside the classical visualization of the factorial plane of variables. Subsequently, a linear regression is employed to observe variable behavior over time. This comprehensive analysis provides insights into the intricate dynamics of viral trends on TikTok, offering a nuanced understanding of their temporal evolution and impact.

TRACKING ARCHIVE'S DATA REUSE IN THE SOCIAL SCIENCES: AN INVESTIGATION

Filippo Accordino[1], Damiana Luzi[2], Fabrizio Pecoraro[3]

[1] *University of Rome La Sapienza* (email: filippo.accordino@uniroma1.it)
[2] *University of San Marino* (email:d.luzzi@unirsm.sm)
[3] *Institute for Research on Population and Social Policies - National Research Council (IRPPS-CNR)* (email: fabrizio.pecoraro@irpps.cnr.it)

Scopus indexed articles metadata, CESSDA Datasets metadata available in CESSDA Data Catalogue Data reuse is a topic of growing importance in research. This is also true in social sciences, where sharing practices and reuse are less common than in other disciplines. The data reuse constitutes a main purpose for the data archives, that consider it also as a key performance indicator. However, identifying the reuse is currently difficult, due the lack of automatic tools and suitable procedures to do it.
The purpose of this study is to track the reuse of datasets stored in social science data archives and cited in scientific publications. The case is focused on two important data archives: GESIS and UK Data Service, both joined to the CESSDA infrastructure.
We identified data reuse by tracking cited data in the scientific publications indexed by Scopus, using references metadata and persistent identifiers. We described most reused data, through their metadata retrieved by CESSDA Data Catalogue and the features of citing publications.
We confirm some critical points in the discovery of data reuse and suggest some solutions to improve it, that involves the whole scientific community.

EMPOWERING SOCIAL SCIENCES: THE ROLE OF DASSI AND FOSSR IN PROMOTING DATA USAGE

Mario Ciampi[1], Massimiliano Saccone[2], Marco Sprocati[3]

[1] *University of Naples Federico II* (email:mario.ciampi@unina.it)
[2] *National Research Council* (email: massimiliano.saccone@cnr.it)
[3] *NIRCRES-National Research Council* (email: marco.sprocati@ircres.cnr.it)

DASSI is the new Italian Service Provider of CESSDA (Consortium of European Social Science Data Archives) ERIC. Its main purposes are the acquisition, preservation and distribution of data for research, following best practices and common European standards. Its activities support and promote the adoption of Open Science and FAIR (Findability, Accessibility, Interoperability, and Reusability) principles in the social sciences, including through training and dissemination activities.
Starting with an introduction to DASSI in the Italian context, the paper will focus on the challenges posed by the availability of new data sources and computational social sciences in terms of data curation and preservation. In this context, the role of DASSI, together with other Research Infrastructures (RIs) such as SHARE, RISIS, GUIDE and GGP, within the FOSSR (Fostering Open Science for Social Research) project will be discussed. The aim of the project is to create a network for harnessing innovation across RIs, with the aim of creating an Italian Open Science Cloud to facilitate remote access to high quality social science data, respecting the FAIR principles and interoperability standards. It aims to foster the use of high quality data among researchers and non-academic users by providing expertise, tools and services for data collection, analysis and harmonisation. Other initiatives include training new researchers, establishing distributed data centres, empowering decision-makers through policy-oriented meetings and creating a sustainable governance model. This comprehensive approach underlines the network's commitment to advancing research on economic and societal change, while ensuring its long-term viability beyond project completion.

ROLE OF THE SOCIAL SCIENCE DATA ARCHIVES IN DEALING WITH PANDEMIC. THE CASE OF THE CESSDA AND COVID19

Yana Leontiyeva[1], Ilona Trtíková[2], Martin Vávra[3]

[1] *Czech Academy of Sciences* (email: yana.leontiyeva@soc.cas.cz)
[2] *CSDA ISCAS* (email: ilona.trtikova@soc.cas.cz)
[3] *CSDA ISCAS* (email: martin.vavra@soc.cas.cz)

In the presentation, we will discuss how European social science data archives (primarily those associated in CESSDA) have been colaborating on immediate solutions and later on tools that can contribute to management of pandemics, such as Covid19. After the COVID19 pandemic broke up, social science data archives have been able to adapt their operations to the pandemic and, in addition, under the coordination of CESSDA, have begun to actively contribute to making data relevant to epidemic management available. Since 2021, CESSDA has been participating with some member archives in the BY-COVID infrastructure project, which is developing tools for making data relevant to the management of the COVID pandemic19 (and management of future pandemics) avaiable under FAIR conditions. The presentation concentrate on various issues connected with CESSDA reaction to pandemic.
There are several challenges CESSDA faces here. One of them is interdisciplinarity (e.g. colaboration with disciplines using different metadata standards, expecially medical sciences and life sciences), data management of new data or sensitive data. Descriptive case study, using documents analysis and interviews analysis have been descripted.

TOWARDS THE CREATION OF A COMPREHENSIVE KNOWLEDGE GRAPH FOR ENABLING SOCIAL SCIENCE RESEARCH

Lorenzo Giammei[1], Misael Mongiovì[2], Andrea Giovanni Nuzzolese[3], Andrea Orazio Spinello[4], Giusy Tuccary[5], Antonio Zinilli[6]

[1] *CNR - Research Institute for Sustainable Economic Growth* (email: lorenzo.giammei@outlook.com)
[2] *University of Catania* (email: misael.mongiovi@unict.it)
[3] *University of Bologna* (email: andrea.nuzzolese2@unibo.it)
[4] *CNR - Research Institute for Sustainable Economic Growth*
[5] *CNR-Research Institute of Cognitive Sciences and Technologies* (email:giusy.tuccari@istc.cnr.it)
[6] *CNR - Research Institute for Sustainable Economic Growth (*email: antonio.zinilli@ircres.cnr.it)

Open science is a rich source for a variety of knowledge intensive tasks such as the analysis, exploration and discovery of research trends and dynamics, such as [1]. A clear example is the discovery of career trajectories and mobility. Unfortunately, open science is still far from delivering research artefacts and data in machine-understandable formats. In the context of the project Fostering Open Science in Social Science Research[1](FOSSR) we are currently experimenting with the creation of a comprehensive semantic knowledge graph (3) (KG) consisting of an ontology and linked open data. Such a KG is built on top of two datasets: (i) the collection of doctoral thesis managed by the Italian National Library of Florence and Rome; and (ii) CercaUniversità[2]. Both datasets are multidisciplinary as they cover all scientific areas.

The construction of the KG is based on an extension of eXtreme Design [2] (XD) that involves: (a) designing an OWL ontology for formalising a shared understanding over the domain that can be user as reference language for modelling data; (b) employing advanced data cleansing and preprocessing techniques to mitigate the issues of noisy and non-standardized data sources; (c) implementing deduplication algorithms and normalisation processes to ensure the integrity and uniformity; (d) tackling the complexity of linking entities across multiple sources by leveraging sophisticated entity resolution techniques based on deep learning; (e) generating linked open data.

The resulting KG harmonises and aggregates data that were previously noisy and non-standardized, e.g. missing or incorrect information, same type of information available into different fields, etc.

The FOSSR project aims at building the Italian Open Cloud in the Social Sciences. Hence, KG is a breakthrough in the realisation of the knowledge layer of such an Open Cloud.

References

(1) Osborne F, Motta E, Mulholland P. Exploring Scholarly Data with Rexplore. In: Salinesi C, Norrie MC, Pastor Ó, editors. Advanced Information Systems Engineering [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013 [cited 2024 Jan 31]. p. 460–77. (Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, et al., editors. Lecture Notes in Computer Science; vol. 7908). Available from: http://link.springer.com/10.1007/978-3-642-41335-3_29

(2) Blomqvist E, Hammar K, Presutti V. Engineering Ontologies with Patterns – The eXtreme Design Methodology. In: Ontology Engineering with Ontology Design Patterns. IOS Press; p. 23–50. (Studies on the Semantic Web; vol. 25: Ontology Engineering with Ontology Design Patterns).

(3) Hogan A, Blomqvist E, Cochez M, D'amato C, Melo GD, Gutierrez C, et al. Knowledge Graphs. ACM Comput Surv. 2022 May 31;54(4):1–37.

IMPROVING POLICY AND TOURISM PLANNING WITH SMART DATA INTEGRATION

Giorgio Garau[1], Giancarlo Onnis[2], Adriano Colosimo[3]

[1] *University of Sassari* (email: giorgio@uniss.it)
[2] *University of Sassari*
[3] *University of Sassari*

In a recent work (1), the foundations are laid for an effective interaction between traditional information and data from unstructured sources. The paper emphasizes, within the context of public policy evaluation, the need to organize information according to the SIS scheme and to use the ADSS as a control panel to verify the coherence of signals from various subsystems. Specifically, in the perspective of enhancing unstructured data or more generally complex data, it is suggested to use the electronic invoice exchange as an example of advanced monitoring, with characteristics of territorial granularity and temporal timeliness.
In the specific context of tourism, it is proposed to use both structured and unstructured sources that collect tourists' feedback regarding their stays, specifically Channel managers and OTAs. The use of such information will enable the profiling of non-seasonal tourists, a segment of particular interest both from the private side (for adapting the offer) and from the public side, which, through this activity, could decide whether and how to invest in deseasoning. This new function, managed through ADSS (prediction of tourist flows by type of tourism) will then be cross-referenced with the macroeconomic subsystem, providing estimates of sustainable tourists derived from the Garau El Meligi model (2). This way, it would be possible to make the interaction between traditional tools of macroeconomics and the opportunities offered by the intelligent processing of data effective, as the data would be valued for its dual utility – for the private sector as a tool to adapt the offer and for the public sector to plan and stimulate the evolution of demand.

References

(1) Garau G., Antolini F., Schirru L., Onnis G. and A. Colosimo, Data and Models: The Role of Statistical Information Systems, Book of Abstracts of the 2. nd Italian Conference on Economic Statistics, Firenze (2024)
(2) Garau G., El Meligi A. K. and D. CARBONI, Perceived crowding and physical distance rules: a national account perspective, National Accounting Review, (2021).

DISSECTING COASTAL AND INLAND TOURISM IN SARDINIA: A STUDY BASED ON ONLINE REVIEWS AND GEOGRAPHIC DICHOTOMY THROUGH NATURAL LANGUAGE PROCESSING

Giulia Contu[1], Cinzia Dessì[2], Carla Massidda[3], Marco Ortu[4]

[1] *University of Cagliari* (email: giulia.contu@unica.it)
[2] *University of Cagliari* (email: cdessi@unica.it)
[3] *University of Cagliari* (email: massidda@unica.it)
[4] *University of Cagliari* (email: marco.ortu@unica.it)

Since the late 1970s, a considerable body of literature on tourism has focused on customer satisfaction rom different perspectives and for different purposes (3). Most satisfaction studies are based on expectation and perception models, cognitive evaluation, congruity and equity models, and perceived overall performance (8). The expectation and perception models proposed by Oliver's (7), which expresses consumer satisfaction as a function of expectation and expectancy disconfirmation, is one of the most used in the definition of tourism satisfaction (TS). Following this model, Truong and Foster (10) have defined the TS as "the difference between expectations and perceived performance, and the "fit" between tourist expectations and host destination attributes" (10). Similar definitions are proposed by other researchers (see for instance, 9, 2). Moreover, same researchers have underlined the necessity to investigate the TS taking into account different aspects. For instance, Yüksel and Yüksel (11) have stated the TS can be evaluated taking into account three different levels: the Product-service level satisfaction, which refers to individual product-service experiences delivered by a single organization in the production chain; the Dimensional level satisfaction, which is obtained summing the satisfactions derived from individual products and services within the given component of tourism; finally the Total satisfaction, which is obtained summing together the individual products-service level satisfactions and dimensional level satisfactions.

Positive results in terms of tourism satisfaction influence the development of a destination and the profitability of private businesses, stimulating tourist expenditure, repeat visits, positive recommendations, and reputation enhancement. Consequently, measuring tourism satisfaction and its determinants becomes crucial for policymakers and managers. Several empirical studies focus on these issues following both qualitative and quantitative approaches. Among these, recently, increasing interest has emerged in measuring customer satisfaction directly or indirectly from online reviews with various scopes and methods (12). This study aims to contribute to this recent line of study by investigating information contained in online customer reviews (e-WOM) in the form of text and scores. The main goal is to understand which is the within-regional variation (Coastal and Inland Tourism) in TS in Sardinia through Online Review Analysis using natural language processing. More in detail, the purpose is to understand if there are differences in topics used by customer reviewers in their online review, and how these differences impact customer satisfaction between coastal and inland geographical areas. At this scope, after investigating each topic's impact on the Sardinian region's customer satisfaction level, we split the dataset into two sub-samples: one related to the destination located in the coastal area, the other in the inland (meant as the central area of Sardinia). Online customer reviews are retrieved from TripAdvisor platforms and related to activities, attractions, and services.

To retrieve and process data from the web, we apply a new method recently proposed by Ortu et al. (8). This method, called TOpic modeling Based Index Assessment through Sentiment (TOBIAS), allows modeling the effects of the topics, moods, and sentiments of the customers' comments describing a phenomenon, such as the perception of the quality of a service, over the level of satisfaction expressed by customers. This method's novelty relies on the combination of natural language processing and causal inference to explain customers' assessment of a phenomenon. TOBIAS is built by combining different techniques and methodologies. Firstly, Sentiment Analysis identifies sentiments, emotions, and moods, and Topic Modeling finds the main relevant topics inside comments. Then, Partial Least Square Path Modeling estimates how they affect an overall rating that summarizes the performance of the analyzed phenomenon.

Since coastal and inland tourism present specific characteristics, we expect to identify differences in topics and impact on satisfaction levels among these two geographical areas by customers. Result are then interpreted. Our contribution is threefold: first, we contribute to tourism literature on customer satisfaction; second, we contribute at the statistical level proposing a new model for analyzing consumer satisfaction; finally, we contribute at the level of policymakers through the interpretation of results by offering strategies to be adopted for the tourism destination.

References

(1) Alkan Y.A. and Kocaman S. (2023), The mediating role of destination satisfaction in the effects of shopping attributes on destination loyalty: the case of Alanya, Journal of Gastronomy, Hospitality and Travel, 6(1), 59- 73.

(2) Chen, C. M., Chen, S. H., & Lee, H. T. (2011) "The destination competitiveness of Kinmen's tourism industry: exploring the interrelationships between tourist perceptions, service performance, customer satisfaction and sustainable tourism", Journal of Sustainable Tourism, Vol 19, No. 2, pp 247-264.

(3) Disegna M., and Osti L. (2016). Tourists' expenditure behaviour: the influence of satisfaction and the dependence of spending categories, Tourism Economics, 22 (1), 5–30 doi: 10.5367/te.2014.0410.

(4) Castro J., Quisimalin M., de Pablos C., Gancino V., Jerez J. (2017), Tourism Marketing: Measuring Tourist Satisfaction", Journal of Service Science and Management, Vol.10 No.3, 2017

(5) Fauzi Sukiman M., Shida I. O., Muhibudin M., Yussof I., Badaruddin M. (2013), Tourist Satisfaction as the Key to Destination Survival in Pahang, Procedia - Social and Behavioral Sciences, 91, pp. 78 – 87

(6) Neal, J.D., and Gursoy, D. (2008), 'A multifaceted analysis of tourism satisfaction', Journal of TravelResearch, Vol 47, pp 53–62.

(7) Oliver, R.L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. Journal of Marketing Research, 17(4), 460-469. doi: 10.2307/3150499.

(8) Ortu M., Frigau L., Contu G. (2022), Topic Based Quality Indexes Assessment Through Sentiment, Computational statistics, DOI: 10.1007/s00180-022-01284-7.

(9) Tribe, J., & Snaith, T. (1998). From SERVQUAL to HOLSAT: Holiday satisfaction in Varadero, Cuba. Tourism Management, 19, 25–34.

(10) Truong, T. H. and Foster, D. (2006) "Using HOLSAT to evaluate tourist satisfaction at destinations: The case of Australian holidaymakers in Vietnam", Tourism Management, Vol 27, No. 5, pp 842-855

(11) Yüksel A., Yüksel F. (2008), Tourist Satisfaction: Definitional and Relational Issues, in the Tourism and Hospitality Industry, Nova Science Publishers, New York, NY, pp. 65-88.

(12) Zhou, K.; Yao, Z. Analysis of Customer Satisfaction in Tourism Services Based on the Kano Model. Systems 2023, 11, 345. https://doi.org/10.3390/systems11070345.

FROM RETROSPECTIVE ANALYSIS TO ANTICIPATION OF TOURISM DEMAND: A NEW SCIENTIFIC APPROACH TO DESTINATION MANAGEMENT

Michela Ciccarelli[1], Fulvio Ilario Giannetti[2], Arianna Testa[3]

[1] *Stockholm University* (email: michela.ciccarelli@lybra.tech)
[2] *Lybra Tech*
[3] *University of Rome Tor Vergata*

This paper aims to analyze the potential of a new Big Data source for the tourism sector based on predictive analytics, and the tools that enable its usability. In the context of the pandemic, the limitations of official data sources have been highlighted by the need to have access to real-time data, capable of providing not only detailed but also prospective. The present study focuses on the Travel Data Lake, an innovative project that centralizes the forecast data of the Zucchetti Group's Booking Engines. Through the dashboard Destination are processed and analyzed an average of thirty million searches for overnight stays per day, which come from a sample of about twenty thousand accommodation facilities Italians. The paper examines three application cases: the reaction of tourism demand to the Ischia floods, the effectiveness of the Unexpected Italy advertising campaign, and the use of predictive data. The results show that real-time data were able to fill an information gap on demand reactions to external events, provide a rapid market response to the effectiveness of a marketing campaign, and enable planning and programming of tourism supply at all levels, from individual hoteliers to land management organizations. In summary, the work demonstrates how the use of real-time data can be a key opportunity for the tourism industry, enabling greater effectiveness in planning and managing tourism offerings.

AIR TRANSPORT AND RATE - SETTING ALGORITHMS

Chiara Tincani

University of Verona (email: chiara.tincani@univr.it)

In 2022 and in the first six months of 2023, not only in the air transport, strong tensions were reported, with significant increases in transport prices and marked fluctuations, with no comprehensible reasons. Although the exact determination of the causes for such behaviour on the part of the airlines has remained unsubstantiated, there is a widespread belief that it depended on the systematic use of highly sophisticated mathematical tools for calculating the charge, with algorithms. The Italian legislator declared his conviction that he had to correct imbalances originated by the epidemic situation and increased by the use of Algorithms. The Law No. 136 of 2023, which converted the Decree - Law No. 104 of 2023, refers all initiatives to the Italian Competition and Market Authority, which "ascertains that the algorithmic coordination of fares practised by airline companies in the aviation sector facilitates, implements or in any case monitors an agreement restrictive of competition, even if pre - existing, or ascertains that the level of prices set through a revenue management system constitutes an abuse of a dominant position". Therefore, on the one hand, the rule has a general object and concerns the entire national territory and, on the other hand, it dwells on “algorithmic coordination”. Thus, there has been a shift from an intervention on the use of algorithms as an intrinsic factor of potential risk to a view of their significance for the preservation of the competitive structure of the market. The subject of territorial continuity appears incidentally, since the Guarantor Authority may consider the fact that illegitimate conduct is "practised on national routes connecting with the islands", "during a period of peak demand linked to seasonality or in conjunction with a state of national emergency", and leads "to a sale price of the ticket or accessory services, in the last week prior to the flight, that is more than two hundred per cent higher than the average fare of the flight". These factors are always relevant in relation to agreements restricting competition or forms of abuse of a dominant position and, therefore, in connection with the distortion of the competitive structure of the market.
Thus, the interventions on the use of algorithms have a declared procedural slant and are linked to the overall assessment of the market structure, with a planned intervention by the Italian Guarantor Authority, intended to promote an orderly flow of flights to the islands and, more generally, competition on all domestic routes. In this perspective, the so - called profiling and the automated analysis of the "type of electronic devices used for bookings" take on importance, in a broader examination of the operators' behaviour.
The same Authority has launched a survey in 2023, aimed at verifying the spread of algorithms in so - called revenue management systems used by airlines, with the dual objective of assessing the possible distortion of competitive dynamics and the possible prejudice for consumers in terms of chances. Moreover, the
Authority has also recently raised questions relating to the determination of air fares, and has not itself identified any infringement, in particular the existence of restrictive agreements.
The issue suggests a broader reflection, with regard to other sectors of tourism; we can think of the hotel where the use of pricing procedures based on user profiling is imagined, especially with regard to the resourses of the customer. Perhaps the investigation can focus on consumer protection, regardless of whether there is a real change in competitive dynamics. The point is to assess whether these pricing mechanisms are unfair commercial practices, with possible recourse to the safeguards provided by the Consumer Code.

THE USE OF BIG DATA IN THE TOURISM SECTOR

Massimo Aria[1], Rosanna Cataldo[2], Maria Gabriella Grassia[3]

[1] *University of Naples Federico II* (email:massimo.aria@unina.it)
[1] *University of Naples Federico II* (email:rosanna.cataldo2@unina.it)
[1] *University of Naples Federico II* (email:mariagabriella.grassia@unina.it)

Today tourism is considered an increasingly strategic sector and, compared to many others, it is very complex, dynamic and constantly evolving. Hence the need to fully understand the phenomenon, quantify it and outline it. There is an increasingly felt need to have up-to-date and timely information, which is at the same time comparable and methodologically valid.
To date, the search for more timely and economical information sources compared to traditional sample surveys is acquiring ever greater importance in tourism statistics. Amid these challenges, more and more new forms of measuring tourism activities are emerging in recent years (3). We are talking about big data, extremely large or complex data sets that can be tracked digitally (2); (1). The advent of big data, in fact, revolutionizes the world of official tourism statistics. Big data for tourism can also provide important information, not only on the collective behaviour of tourists, but also on the relationship between places, objects and people. They can open up new analysis perspectives with a greater level of detail, such as the temporal dimension and timely monitoring of data on the territory.
This work provides an overview of different sources of big data and their potential relevance in tourism statistics. The opportunities and risks that the use of new sources can create will also be presented.

References

(1) Mariani, M.M., Baggio, R., Fuchs, M. and Hopken, W. (2018), Business intelligence and big data in hospitality and tourism: a systematic literature review, International Journal of Contemporary Hospitality Management, 30 (12), pp. 3514-3554.
(2) Onder, I., Koerbitz, W. and Hubmann-Haidvogel, A. (2016), Tracing tourists by their digital footprints: the case of Austria, Journal of Travel Research, 55 (5), pp. 566-573.
(3) Volo, S. (2018), Tourism data sources: from official statistics to big data, in Cooper, C., Volo, S., Gartner, W.C. and Scott, N. (Eds), The SAGE Handbook of Tourism Management, SAGE, London, pp. 193-201.

DEMOGRAPHIC DYNAMICS IN TOURISM: GLOBAL TREND AND SUSTAINABLE FUTURE

Elisa Cisotto

Free University of Bozen (email:elisa.cisotto@unibz.it)

In the context of the transition towards sustainable tourism, acquiring an in-depth understanding of the central role of demography becomes essential. Demographic changes, albeit gradual, shape the workforce structure, influence available income levels, define the contours of the pension system, and reflect changes in health, daily activities, and consumption choices across different generations. Demography, as an analytical tool, offers the chance to observe the current social structure, predict future dynamics and assess the sustainability of emerging trends.

Within the specific context of the demography of tourism, this study aims to explore the main demographic trends and analyze their impact on the dynamics of tourist demand, service offerings, and human resource management. Using official national and international data sources, the presentation highlights how demographic changes can represent both opportunities and risks for the future of tourism. Among the key evolutions, some major determinants shaping leisure time and tourism include the overall increase in life expectancy and healthy life expectancy, the shrinking size of households, and the emergence of new family structures, along with the growing proportion of individuals with a high level of education and changes in migration patterns.

Finally, based on the emerging data and key theories guiding tourism demography, the presentation outlines some potential implications of ongoing demographic changes for the touristic sector. It emphasizes relevant application areas and underscores the need for flexible, market-oriented strategies to maximize attractiveness and satisfaction across various market segments.

BREAKING DOWN BARRIERS TO TOURISM FOR PEOPLE WITH DISABILITIES: THE ROLE OF SOCIAL CAPITAL

Massimiliano Agovino[1], Katia Marchesano[2]

[1] *University of Naples Parthenope* (email: massimiliano.agovino@uniparthenope.it)
[2] *University of Naples Parthenope* (email: katia.marchesano@uniparthenope.it)

People with disabilities represent a significant and undervalued niche of the tourism industry, which is bound to increase in the coming years, due to increasing life expectancy and population ageing. Disability is indeed a highly age-related phenomenon, whose extent increases more than proportionally with age. Approximately 15% of the world's population features at least one disability. In Italy, about 5.2% of the population faces severe limitations in carrying out daily activities, while 16.4% faces moderate limitations (ISTAT, 2021).
Tourism participation opportunities of these consumers depend crucially on the existence of tourism goods and services that meet their needs in a destination. These ensure that transport, accommodation and attractions, such as larger rooms, adequate lifts, equipped bathrooms and specific staff training. The lack of these tourism goods and services raises barriers to tourism participation for people with disabilities (i.e., environmental, economic and informational barriers – from now on EEI barriers). These EEI barriers affect their choice on the tourism good by inducing a self-selection problem. This problem may generate a biased measure of their probability to participate in tourism. Starting from this consideration, the aim of the study is twofold. Firstly, the theoretical impact of barriers on tourism choices of a person with disabilities and how their existence can generate a self-selection problem are shown. Secondly, the role of social capital in reducing those barriers is investigated. To verify and address these issues, a Heckman selection model is employed using microdata on the Italian population. The results show the importance of social capital, via bridging and linking networks, in reducing the barriers to tourism participation.

GEOGRAPHY AND TOURISM

Dionisia Russo Krauss[1], Maria Ronza[2]

[1] *University of Naples Federico II* (email: dionisia.russokrauss@unina.it)
[2] *University of Naples Federico II* (email: maria.ronza@unina.it)

Although tourism can undoubtedly be considered a geographical phenomenon, research in the discipline only emerged in the wake of the evolution of systemic approaches in the social sciences in the Seventies-Eighties, towards broader explanatory investigations. These studies increasingly focused on the territorial organization of societies, as well as on environments, landscapes, cultures, and the material and immaterial construction of leisure places, narratives, and tourist imaginaries. This shift led to reflections on the spatial impact and consequences of tourism activities. With changes in research methods and the enrichment of a vast and articulated production, the Geography of Tourism has not only seen the emergence of new themes but also the establishment of a new approach. Particularly, the possibility to integrate cartographic sources and statistical data, field-collected data, and those derived from satellite or aerial imagery, effectively undermined the rigid dichotomy of research methodologies (idiographic approach and inductive method/nomothetic approach and deductive method) in geographic research in general and, consequently, in tourism studies as well.
Initially considered merely as a technical-informatic support, GIS (Geographic Information Systems) has become a fundamental skill among the new generations of geographers, aiming to promote territorial studies with a project-oriented dimension focused on sustainability, circular economy, enhancement of cultural and environmental heritage, and eco-compatible use of local resources. After considering the evolution of the discipline's relationship with tourism and reviewing the main scientific paradigms, this contribution focuses precisely on the impact of Geographic Information Technologies, examining some types of GIS applications – and the related investigation methods of the tourism/territory relationship – that have increasingly spread since the mid-Nineties. These include the transition from statistical reports to digital cartography in the analysis of flows and receptivity, the creation of layers (informational strata), and their overlapping for the spatiotemporal analysis of a tourist system, and the development of projects (with WebGIS) for the construction of cultural itineraries and towards a smart tourism perspective. These different methods offer ways to read, interpret, and promote places, with Geography continuing to prioritize attention to how the phenomena under investigation interfere with space or create new spatialities compared to other social sciences.

A DIRICHLET-MUL.NOMIAL MIXTURE MODEL FOR 30 YEARS OF SCHOLARLY PAPERS IN STA.S.CS ON ARXIV

Massimo Bilancia[1], Rade Dačević[2]

[1] *University of Bari Aldo Moro* (email: massimo.bilancia@uniba.it)
[2] *EY Business & Technology SoluCon*

In this paper, we used Bayesian natural language processing methods to collect and analyze a large corpus of 111,411 eprints submi<ed to the arXiv repository between 1994 and 2022 under categories of the StaCsCcs group (the first classificaCon level of eprints on arXiv). We aim to invesCgate the extent to which Machine Learning has contributed to changing staCsCcs as a discipline, i.e. whether the advent of these techniques has brought about a genuine paradigm shiM, replacing old problems with new ones, or whether the revoluCon should actually be sought elsewhere. Science is a metastable state. A metastable state describes a phase in which an energy barrier must be overcome before this phase can be transformed into a phase with lower free energy. Using this metaphor, we can describe the progress of a scienCfic discipline as the accumulaCon of knowledge that makes it possible to overcome a conceptual or computaConal barrier that has prevented the soluCon of certain problems and to reach a state in which the effort required to solve these problems becomes less because of the new knowledge. Old problems are not solved, but simply set aside, and new quesCons become interesCng for the majority of researchers in a scienCfic community. Regardless of whether one agrees or disagrees with the view of science based on these paradigm shiMs, we can conclude, based on the textual analysis of the published preprints, that the only real paradigm shiM that has taken place so far for staCsCcs as a scienCfic discipline in its enCrety is the Bayesian revoluCon that started in the early 1990s.

RELEVANCE OF SDGS INDICATORS IN SUSTAINABLE TOURISM AND DEMOGRAPHIC TRENDS

Najada Firza[1], Angela M. D'Uggento[2], Corrado Crocetta[3]

[1] *University of Bari Aldo Moro* (email: najada.firza@uniba.it)
[2] *University of Bari Aldo Moro* (email: angelamaria.duggento@uniba.it)
[3] *University of Bari Aldo Moro* (email: corrado.crocetta@uniba.it)

The health of a region and its appeal to tourists are primarily measured through demographic indicators. Structural issues such as low birth rates, high aging populations, and the onset of a mature demographic phase affect most advanced economies. Our research provides indicators for evaluating the sustainability of the tourism sector, which are linked to demographic indicators. The study analysed the relationships between the tourism industry and demographic indicators, with a focus on the indicators of goals 9, 11 and 12 of the SDGs. The aim was to highlight the positive economic opportunities that the tourism industry creates at a regional level in Italy.

ASSESSING ITALIAN LEARNING GAPS WITH INVALSI DATA VIA SMALL AREA ESTIMATION

Diego Battagliese[1], Mario Intini[2], Alessio Pollice[3], Angela S. Bergantino[4]

[1] *University of Bari Aldo Moro* (email: diego.battagliese@uniba.it)
[1] *University of Bari Aldo Moro* (email: mario.intini@uniba.it)
[1] *University of Bari Aldo Moro* (email: alessio.pollice@uniba.it)
[1] *University of Bari Aldo Moro* (email: angelastefania.bergantino@uniba.it)

Nowadays, the problem of taking actions in areas showing particular deficiencies has become more and more important for the government agencies. Official statistics and agencies in general require reliable estimates of socio-economic quantities of interest, such as poverty or unemployment, in order to allow political and economic decision makers to adopt appropriate interventions. However, some subpopulations may be difficult to observe and the lack of suitable data in areas with limited number of units has sharpened this problem.

Small area estimation techniques try to overcome this issue. In particular, area-level models exploit the information coming from a direct estimator of small area means or totals. Direct estimators arise from a sample design to which a sampling error is typically associated. Existing surveys are not always planned for disaggregated territorial levels, thus data may be limited, leading to small precision of the estimates. The direct information is then combined with a model-based synthetic estimator in order to have more robust results. Within the class of area-level models, the Fay-Herriot model is the most popular.

We make use of INVALSI test scores as a direct estimator in order to asses learning gaps across Italy. In addition, we take advantage of some auxiliary variables at the specific area level. The general Fay-Herriot estimator is a convex linear combination of direct and synthetic estimators, with a shrinkage factor regulating the weight associated to them.

We propose a Bayesian hierarchical model in order to estimate the model parameters. We also propose a strategy to scale direct estimators to smaller territorial levels with an additional level of hierarchy. The additional layer allows to smooth the variance estimation and have more accurate estimates.

Acknowledgement

This study was funded by the European Union – *NextGenerationEU*, in the framework of the *GRINS – Growing Resilient, INclusive and Sustainable* project (GRINS PE00000018 – CUP H93C22000650001). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

IDENTIFICATION OF PERSPECTIVE DATA FROM THE EUROPEAN SOCIAL SURVEY (ESS) FOR THE DEVELOPMENT OF SMART CITIES AND SMART PEOPLE

Francesco D. D'Ovidio[1], Silvia D'Ovidio[2], Antonella Nannavecchia[3]

[1] *University of Bari Aldo Moro* (email: francescodomenico.dovidio@uniba.it)
[2] *Freelance Scholar, Bruxelles (Belgie).and IARC-UNICART Str*. (email: silvia.dovidio@sciencespo.fr)
[3] *Freelance Scholar, and Tourism entrepreneur* (email: nannavecchia.a@gmail.com)

The interaction of people with social networks, smartphones and other technologies produces a significant amount of data. Specific technologies, such as the *Internet of Things* and *Big Data Analysis*, are essential tools for transforming traditional cities into Smart Cities. Six recognised operational areas are pillars for the development of smart cities: smart economy, people, governance, mobility, living and environment. Models that combine these operational areas are useful to reveal the level of smartness of a city.

European governments acknowledge the significance of utilising tools to arrange urban services according to citizens' needs, mobility, and environment. Along with data automatically provided by the devices people use daily, information can also be collected through sample surveys to obtain citizens' opinions. When combined with technological data, preliminary operations are required to identify the subjective information that interacts the most with objective data.

This article utilises data from the 2019 European Social Survey (ESS), a cross-national survey conducted for academic research purposes, to test a valuable methodology. The primary objectives of the survey are to monitor and interpret changes in public attitudes and values across Europe, improve cross-national survey measurement methods, and develop a set of social indicators.

Round 8 of the 2019 ESS covered 23 countries, including EU member states, Switzerland, Israel, and the Russian Federation. Rigorous methodologies were used to gather information and opinions from 44,387 European citizens, resulting in over 400 variables. The main aim is to identify the variables that are important for joint analyses with other data, such as regional economic indicators, in this stream of information.

It is worth noting that statistical inference methods are not applicable to Large and Big Data. Therefore, we tested multivariate statistical methods, including Categorical Principal Component Analysis (CatPCA) and Artificial Neural Network Analysis, specifically Kohonen's Self Organizing Maps (SOM).

THE MACHINE LEARNING CONTROL METHOD FOR COUNTERFACTUAL FORECASTING

Augusto Cerqua[1], Marco Letta[2], Fiammetta Menchetti[3]

[1] *University of Florence*
[2] *University of Florence*
[3] *University of Florence* (email: fiammetta.menchetti@unifi.it)

Without a credible control group, the most widespread methodologies for estimating causal effects cannot be applied. To fill this gap, we propose the Machine Learning Control Method (MLCM), a new approach for causal panel analysis based on counterfactual forecasting with machine learning. The MLCM estimates policy-relevant causal parameters in short- and long-panel settings without relying on untreated units. We formalize identification in the potential outcomes framework and then provide estimation based on supervised machine learning algorithms. To illustrate the advantages of our estimator, we present simulation evidence and an empirical application on the impact of the COVID-19 crisis on educational inequality in Italy. We implement the proposed method in the companion R package MachineControl.

A SIMULATION STUDY TO COMPARE MARMOT ADJUSTMENT AND TEMPLATE MATCHING IN A MULTIPLE TREATMENT FRAMEWORK

Pietro Belloni[1], Alberto Calore[2], Margherita Silan[3]

[1] *University of Padua* (email: pietro.belloni@unipd.it)
[1] *University of Padua* (email: alberto.calore@unipd.it)
[1] *University of Padua* (email: margherita.silan@unipd.it)

The estimation of a causal effect in a multi-treatment framework needs the use of specific statistical tools, especially when the number of treatments considered is particularly large. Case studies involving a huge number of treatments are quite rare in scientific literature and leverage mainly on two methods: Matching on Poset based Average Rank for Multiple Treatments (MARMoT) and Template Matching. The aim of this work consists in the comparison of these two techniques through a simulation study that involves different scenarios. We built those artificial scenarios varying multiple aspects, such as the number of treatments (50, 250, 500), the presence of rare treatments and the presence of rare confounders. Moreover, we implemented small technical changes to the two techniques to improve their performance within the different scenarios. Our final goal is to empirically establish in which setting each of the two techniques perform better than the other. These two methods are compared also with real data coming from Medicare database to compare 41 medical facilities on their performance on elderly patients undergoing cardiac surgeries. The conclusion of this study may provide valuable guidelines for the selection and implementation of statistical approaches in addressing self-selection bias in multi-treatment observational studies.

COMBINING A FINITE MIXTURE APPROACH WITH PROPENSITY SCORE TO MEASURE THE IMPACT OF SOCIAL TIES ON OLDER PEOPLE'S DIGITALIZATION

Dalila Failli[1], Bruno Arpino[2]

[1] *University of Florence* (email: dalila.failli@unifi.it)
[2] *University of Florence* (email: bruno.arpino@unifi.it)

Age digital divide is defined as the gap between older adults and younger individuals in accessing or using different digital technologies. Even if this phenomenon has narrowed in recent decades, older people still show a lower propensity to use the Internet and adopt new technologies. Results from the literature show that social ties with family and friends may have a great influence on the digital inclusion of older adults. To understand the causal link between intergenerational ties and the reduction of age digital divide, we propose to combine data from the European Social Survey (ESS) and the Survey of Health, Aging and Retirement in Europe (SHARE); the former contains information on the digitalization level of older adults, while the latter provides details on the individuals' network of family contacts and friendships. An exact matching of these data sets is performed on the basis of the individuals' demographic characteristics, such as age, health, gender, education, income, country, and employment status. Under the strong ignorability assumption, we propose to remove covariate imbalance between treatment groups through propensity score matching techniques. A hierarchical finite mixture model is then applied to the matched data set, where exposure indicators are added as predictors of latent class membership probabilities to inform about the causal effect of interest under strong ignorability.

INTEGRATION AND PREDICTIVE MODELING OF MICRODATA IN TOURISM

Fabrizio Antolini[1], Samuele Cesarini[2], Ivan Terraglia[3]

[1] *University of Teramo* (email: fantolini@unite.it)
[2] *University of Teramo* (email: scesarini@unite.it)
[3] *University of Teramo* (email: iterraglia@unite.it)

This study outlines the development of an integrated data system specifically tailored for the Italian tourism sector, achieved through the combination of microdata from surveys conducted by the Bank of Italy and ISTAT. The primary aim is to provide a comprehensive analysis of both national and international tourism aspects, thereby enabling exploratory analyses and the construction of useful indicators. This is achieved by integrating machine learning predictive models able to maximizing the potential of microdata and yielding better results in terms of accuracy (1). The methodology employed involved building a SQL database through Extraction, Transformation, and Load (ETL) processes, coupled with the application of advanced analytical techniques and predictive models. The results of this study are two-fold, theoretically, it aims to establish a coherent integrated system reflecting the complex and multi-segmented nature of tourism phenomena. Practically, it seeks to provide an effective tool for public policy makers and industry stakeholders. The use of decision tree algorithm not only enhances the understanding of tourism patterns but also lays a solid foundation for more informed and strategic decision-making in the tourism field (1).
Furthermore, the innovative approach of combining various types of data allows for a more detailed view of tourism (2), considering variables such as spending trends, traveler preferences, and economic impact. This method of data analysis aims to contribute to the evolution of monitoring and managing the Italian tourism sector, potentially offering benefits at both national and international levels.

References

(1) Antolini, F., Cesarini, S., Simonetti, B. (2024). Italian tourist expenses' determining factors: a machine learning approach. Qual Quant. https://doi.org/10.1007/s11135-024-01832-x..
(2) Antolini, F., Terraglia I., Cesarini, S. (2023). A composite-indicator approach for assessing sustainable tourism of Italian regions in XV Riunione Scientifica SISTUR, Messina - Taormina.

IMPACTS OF COHESION FUNDS ON LOCAL TOURISM. COUNTERFACTUAL ANALYSIS AND MACHINE LEARNING APPROACHES

Gianluca Monturano

Univesity of Modena and Reggio Emilia (email: gianluca.monturano@unimore.it)

This study, exploiting the counterfactual econometric methodology, assesses the impact of the Territorial Cohesion Funds 2014-2020, which follow a placebased local development approach, on the performance of the Italian tourism sector. In particular, through the analysis of ISTAT data relating to tourist presences and arrivals in municipalities, the influence of funding on tourist flows is quantified.
The Italian spatial heterogeneity, which influences development, has also been taken into account, repeating the estimates with respect to recent municipal classifications that identify the "prevailing tourist classes" and the inland areas.
In addition, Machine Learning algorithms were used to predict future tourist flows. This approach makes it possible to assess not only past impacts, but also to provide predictions based on empirical analyses. The results of this study contribute to a better understanding of local tourism dynamics and the effectiveness of territorial development policies in support of the tourism sector.
Empirical evidence and advanced forecasts provide valuable information for the planning and management of resources in the regions covered by the Territorial Cohesion Funds.

ANALYZING THE TOURISM BEHAVIOR PATTERNS IN SARDINIA. A MARKOV CHAIN APPROACH TO INVETIGATE THE MOVEMENTS OF THE TOURIST INSIDE THE ISLAND

Giulia Contu[1], Marco Ortu[2], Andrea Carta[3], Luca Frigau[4]

[1] *University of Cagliari* (email: giulia.contu@unica.it)
[2] *University of Cagliari* (email: marco.ortu@unica.it)
[3] *University of Cagliari* (email: andrea.carta88@unica.it)
[4] *University of Cagliari* (email: frigau@unica.it)

Studying the movement of tourists inside a destination can be useful to comprehend the spatial temporal behavior of tourists, and to investigate the existence of specific paths followed by the tourists. Moreover, comprehending which elements can impact on the movement can support the government to identify the existence of a specific target and, consequently, create adequate services to support the tourist during the time spend in a destination.
This paper aims to investigate the movement of tourists inside Sardinian Island using the data of tourist information points. The Sardinian tourism office managed by the regional government has introduced a new system to record information related to the tourists that visit the tourist information points.
The recorded information allows knowing for each tourist that visits the tourist information point: the nationality, the transport main used to arrive in Sardinia and used to move around the different places, and the typology of accommodation. Additionally, the tourists communicate the destination visited before to come in Tourist information point and the next destination. We use the recorded information to identify firstly the main paths chosen by the tourists in Sardinia, secondly to estimate the probability that a tourist chooses a specific tourist destination given he/she has been to another destination before; thirdly to comprehend which elements can impact on the choice of the next destination.
Two different methodologies have been used. The first is the complex network approach useful to investigate the path between the destinations and to identify the cities most visited inside the Island.
The second is the Markov chain approach to investigate the probability that a tourist chooses a specific destination aJer he/she has been at another destination.
The first results show the presence of specific tourist paths inside the Island and the main movements is recorded in the south of the Island.

STUDENT COMMUTING IN ITALY TRAJECTORIES: A STUDY ON TWO PROVINCES OF THE CENTRE-SOUTH AREAS

Oliviero Casacchia[1], Cecilia Reynaud[2], Salvatore Stozza[3], Enrico Tucci[4]

[1] *University of Rome La Sapienza* (email: oliviero.casacchia@uniroma1.it)
[2] *Roma Tre University* (email: ceci@uniroma3.it)
[3] *University of Naples Federico II* (email:Salvatore.strozza@unina.it)
[4] *ISTAT* (email: tucci@istat.it)

In Italy, one-third of the systematic daily moves in the country are produced by students. Many research concern the mobility of the employed, scarce are those devoted to students mobility (1). The paper is devoted to analyse the characteristics of student commuter mobility of two contiguous provinces (Frosinone and Caserta), concerning inter-municipal daily moves observed at the 1991-2011 censuses. The methodologies used - additive log-linear and gravity models (2) - permit to synthesize the large amount of information contained in the origin-destination matrix of the commuting flows (nearly 60 thousand cells considering the three censuses) into few parameters which permit to distinguish the propensity of the origin area to produce out-flows and that of the destination area in which commuters study to attract flows (3). All the developed models control for the effect of contiguity and spatial structure of the selected areas. After a brief descriptive analysis of the origin-destination matrices conducted using some traditional indicators (self-containment, prevalence, attraction/repulsion effect), enlarged gravity model taking into account the spatial structure of the
selected areas are built. Log-linear additive model (in various versions) is developed to study the interaction between origin and destination areas and to test that structure of migration is to be invariant with respect to time and sex.The parameters obtained with the two applications described above permit to synthetized the main component of the commuting behaviour of the students in the two selected areas. Some comparison with results obtained considering work commuters are also discussed in the paper.
Gravity models fits fairly well the origin/destination matrix of the students who go to the study place each day. Log-linear additive models permit to easily put in comparison various matrices describing the spatial structure of the commuters in various years and in the two selected provinces.

References

(1) Sbianchi, G., Pascucci, S. and Vitiello C. (2017). Il pendolarismo scolastico: un'analisi nell'area metropolitana di Roma Capitale, rivista Italiana di Economia, Demografia e Statistica, Vol. 71, No. 1, pp. 5-16.
(2) Little J and Raymer J (2013). Log-linear models of migration flows. In Tools for demographic estimation, Moultrie T, Dorrington R, Hill A, Hill K, Timaeus I and Zaba B, Eds., International Union for the Scientific Study of Population, Paris, pp. 403-419.
(3) Casacchia O, Reynaud C, Strozza S and Tucci E (2022). Internal migration patterns of foreign citizens in Italy, International Migration, Vol. 60, No. 5, pp. 183-197.

THE PROPENSITY OF STUDENTS TO SUSTAINABLE MOBILITY

Simona Balzano[1], Houyem Demni[2], Luisa Natale[3], Edoardo Pascucci[4], Giovanni C. Porzio[5]

[1] *University of Cassino* (email: s.balzano@unicas.it)
[2] *University of Cassino* (email: houyem.demni@unicas.it)
[3] *University of Cassino* (email: natale@unicas.it)
[4] *University of Cassino* (email: edoardo.pascucci@unicas.it)
[5] *University of Cassino* (email: porzio@unicas.it)

In this work we present the first results from a survey on the propensity of students to adopt sustainable mobility behaviours in the metropolitan area of Rome, including the measurement of the level of risk perceived by vulnerable road users (pedestrians and bikers) in urban routes. The survey is aimed at the population of 18-35 aged people in the metropolitan city of Rome, who commute for study reasons at least once a week. Data collection is managed through a CATI system. It involves a sample of 1000 students, balanced by the means of transportation they mainly use to reach their study place. The structure of the questionnaire was inspired by the well-known Technology Acceptance Model (TAM) (1), based on a latent variable model, adapted and integrated with some typical constructs borrowed from the main literature on the consumer propensity to product purchase (2, 3), i.e. perceived risk, perceived usefulness, ease of use, push and pull factors, intention to use, use behaviour, where the "use" refers to the adoption of sustainable behaviours in urban mobility routes. The research is part of a wider study on social sustainability in transportation aiming at define guidelines and recommendations for agencies, policymakers, urban planners and other stakeholders involved in the design and forthcoming construction of the Tecnnopole building of Sapienza University of Rome, that will be located in the Pietralata area of Rome. The study aims to contribute valuable insights to help urban planners and other players to create an efficient mobility systems in the whole area.
Among the main expected results we underline the definition of users' profiles; the identification of the different degrees and dimensions of risk perception by the vulnerable users; the identification of the main causes of insecurity arising from routes characteristics, weather conditions, traffic conditions, etc.; the identification of the motivations that encourage or discourage users to adopt sustainable mobility behaviours.

References

(1) Davis, F.D., Bagozzi, R.P., Warshaw, P.R. (1989). User acceptance of computer technology: a comparison of two theoretical models, *Management Science*, Vol. **35**, No. 8, pp. 982-1003
(2) Kaplan L.B., Szybillo G.J., Jacoby J. (1974). Components of perceived risk in product purchase: a cross-validation. *Journal of Applied Psychology*, **59**(3), 287-291, DOI: 10.1037/h0036657.
(3) Marton G, Monzani D, Vergani L, Pizzoli SFM, Pravettoni G. (2023). How to Measure Propensity to Take Risks in the Italian Context: The Italian Validation of the Risk Propensity Scale, *Psychological Reports*, **126**(2):1003-1017. doi: 10.1177/00332941211054777. PMID: 34879777.

AN INDEX OF THE STUDENT MOBILITY FLOW BETWEEN UNIVERSITIES OF DIFFERENT SIZES

Giuseppe Giordano[1], Ilaria Primerano[2]

[1] *University of Salerno* (email: ggiordano@unisa.it)
[2] *National Research Council of Italy - Institute for Research on Population and Social Policies (CNR-IRPPS)* (email: ilaria.primerano@cnr.it)

The issue of undergraduate students' mobility among different Universities during their careers is addressed considering the specific impact that these flows have on origin and destination Universities in terms of their retention ability and attractiveness. This phenomenon has been defined with the expression "students churn risk", referring to the possibility of students to change Universities as a risk for the University of first enrolment, i.e., as a loss in terms of student population numbers and, vice versa, as an opportunity for the University of second enrolment.
Moving from this framework, the aim of this contribution is to propose an index of students' mobility flows based on retention ability and students churn decisions.
At this aim, also considering the characteristics of the Universities in terms of their size, we propose a normalize weighting system for incoming and outcoming flows. Thus, Social Network and Multidimensional Data Analyses are used to visualize and explore roles and positions of specific Universities in the whole network. In the framework of network data analysis, these flows can be arranged into a weighted, directed bipartite network defined by: i) a set of node-origins representing the initial sources (University of first enrolment), ii) a set of node-destinations (University of second enrolment), and iii) a set of arcs joining the origins and destinations, weighted by the occurrence of students moving between any pairs of nodes belonging to two sets.
Some case studies analysed at different geographical level (regional and national) are discussed considering the choices of Italian students to change University when moving from the first to the second year of a bachelor's degree program.

DSSR 2024
Naples, 25th-27th March

Data Science & Social Research
4th International conference

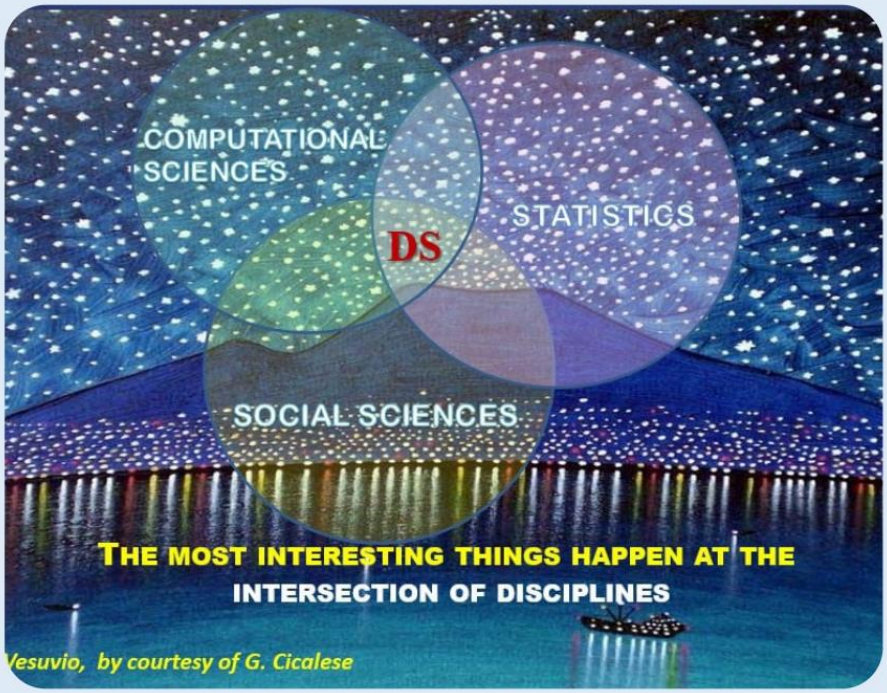
Abstract book

DSSR 2024

Data Science & Social Research

4th International conference

Naples, 25th-27th MARCH



Organised by
Department of Social Sciences (DISS) and
Department of Economics and Statistics (DISES)
University of Naples Federico II

PRIN 2022- The Future of Sustainability – P2022B3NFH – H53D23009470001-E53D23019030001