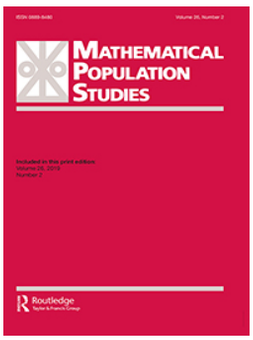

Routledge
Taylor & Francis Group

Mathematical Population Studies

An International Journal of Mathematical Demography

ISSN: 0889-8480 (Print) 1547-724X (Online) Journal homepage: https://www.tandfonline.com/loi/gmps20

Methods for big data in social sciences

Enrica Amaturo & Biagio Aragona

To cite this article: Enrica Amaturo & Biagio Aragona (2019) Methods for big data in social sciences, Mathematical Population Studies, 26:2, 65-68, DOI: 10.1080/08898480.2019.1597577

To link to this article: https://doi.org/10.1080/08898480.2019.1597577

MATHEMATICAL POPULATION STUDIES
2019, VOL. 26, NO. 2, 65–68
https://doi.org/10.1080/08898480.2019.1597577

Routledge
Taylor & Francis Group

Methods for big data in social sciences

Enrica Amaturo and Biagio Aragona

Department of Social Sciences, University of Naples Federico II, Naples, Italy

1. Different forms of digital data

The diffusion of digital technologies and social networks has multiplied the forms of digital data that can be employed for social research.

The main two forms are native digital data, which are produced in social networks, search engines, or blogging, and digitized data, which are analog data transformed into digital (Rogers, 2013).

Big data are originally produced in the Internet. They allow for analyzing behaviors without interfering with individuals (Webb et al., 1966). An example is the data used in web platforms analytics, such as *Google Correlate*, whose purpose is to reveal the co-occurrences associated with a keyword searched through the *Google* search engine. This tool helped to predict the flu epidemic in the US, well before the US Centre for Disease Control and Prevention (Ginsberg et al., 2009). This example demonstrates that digital web platforms enable innovations in data analysis. Another example of native digital data is the data voluntarily uploaded on social networks, blogs, and websites. These are mainly textual or visual (images and videos), often unstructured. A third example is transactional data and the Internet of things. Transactions made through digital devices, such as smart-phones, scanners, tablets, and cards with chips (credit cards, shopping cards) produce data with some structure. These data comprise metadata (date, time, duration, or expenditures) associated with transactions. The objects connected to the Internet (the Internet of things), such as sensors for health monitoring, house automation, and driving aid, usually produce structured data, which can be organized and analyzed.

Digitized data previously existed in analog form, for example images, videos, and scanned or digitally photographed documents uploaded on the web, such as museum collections or libraries available on-line. Digital humanities have converted this material into digital form. Another example is the surveys assisted by computers, where the data are inserted into digital databases. Web surveys now are conducted through the Internet (by e-mail) (Amaturo and Aragona, 2016), and allow for reaching a large sample with a small budget.

CONTACT Enrica Amaturo amaturo@unina.it Department of Social Sciences, University of Naples Federico II, vico Monte di Pietà, 1, Naples, Italy.

Digital data however require adequate methods. They do not necessary demand computational techniques, but specific skills. For example, machine learning, sentiment analysis, or social network analysis are rooted in content analysis, agent-based modeling, or network analysis.

2. Digital data require specific methods

The abundance and granularity of social media data have empowered and transformed network analysis. This latter technique has been used in sociology (Latour, 2005; Scott, 2012) and can be traced back to the sociometric work of Moreno (1934), who mapped out likes and dislikes among members of small social groups, such as school classes and sport teams. Marres (2017) notes that since Moreno's work, network analysis has been developed along "mathematical (graph theory), quantitative (social network analysis), and radically empiricist (actor-network theory)" (92). Social network data allow empiricism in graph theory (Newman et al., 2007; Lazer et al., 2009) and a shift from modeling networks to the analysis of real-time network dynamics (Escobar et al., 2017). Small group studies have been replaced by the analysis of social media platforms, in order to study network dynamics on a large scale (Rieder, 2013). The structure, patterns, and trends of data objects and their relations are often systematically visualized.

Scientometrics (De Solla Price, 1978) consists of the quantitative analysis of literature (bibliometrics) run on digital bibliographical data infrastructures such as the *ISI Web of Science, Elsevier Scopus, Google Scholar*, or on digital archives. Network analysis helps to map digital references to books and articles according to citations, mentions, time, subjects, and other variables.

Digital data enhanced also content analysis (Berelson, 1952; Amaturo and Punziano, 2013). Herring (2009) claims that digital content analysis has distinctive innovative features, such as the possibility to visualize words and their links, and to analyze them in real time. Sentiment analysis is devised for analysis of the human language on the web. It uses semantics and taxonomies to recognize and extract patterns from posts, tweets, comments, and web documents. Its purpose is to characterize opinions about an issue. It is based on a thesaurus of sentiments, reflected by words which from the context hold either positive or negative meanings (for example "good" may score +2 and "terrible" −3). The sum of scores of all the words contained in the document measures the mood with regard to the topic.

Machine learning is a branch of artificial intelligence. It has been developed for exploiting big data. Very large datasets can be analyzed timely only by algorithms. Machine-learning algorithms are automated and "learn" from the data. It is used to recognize patterns in datasets and to construct models of these patterns (Han et al., 2011). Supervised machine learning uses training data to develop learning processes, which consists in matching inputs

with certain outputs. Unsupervised learning spots itself patterns and structures in the data, without preliminary training data. Machine learning is used for data mining (Manyika et al., 2011) and for detecting, classifying, and segmenting meaningful relationships between variables. Data mining may employ neural networks, decision trees, and statistical (parametric or not) methods.

Public administrations are also developing digital data. Statistical offices are now financing open-data infrastructures and inserting big data in the production of official statistics. This raises the question of the validity of digital data in reflecting social processes and their use in conducting public policies.

3. In this special issue "Methods for big data in social sciences"

Luis Martinez-Uribe shows that collections prepared by libraries can be used as big data. He uses network coincidence analysis, a method for combining co-incidence and social network analyses, on more than three million records, which represent 800,000 person names and 300,000 subject headings of the British National bibliography.

Alessandra Righi exploits social network data to measure migration flows, the integration of migrants in destination countries, and public opinion toward migrants. She expresses the need for data access and partnership with data providers to overcome legal obstacles. She explains how *Twitter* data can be customized for measuring the sentiment of Italian-speaking users against migration.

Angela Chieppa, Gerardo Gallo, Valeria Tomeo, Francesco Borrelli, and Stefania Di Domenico present a data infrastructure from the Italian national institute of statistics (*Istat*) associating official population registers with other subject-specific administrative registers. They use machine learning and the knowledge discovery process in order to identify patterns in data. Their technique helps produce accurate population counts. They mention the difficulties encountered in reaching subpopulations.

Biagio Aragona and Rosanna De Rosa review studies where digital data can facilitate public policies. They show the risk of collecting data with techniques unknown by stakeholders. They question the validity of big data and plead for integrating big data with surveys and censuses.

Digital data do not replace surveys. Maria Michela Dickson, Anton Grafström, Diego Giuliani, and Giuseppe Espa simulate sampling schemes in establishment surveys. They propose a sampling procedure based on spatial sampling to be employed in establishment surveys. Stratified sampling has mostly been used in surveys on businesses. The authors overcome the problems of high stratification that may compromise implementation of a sample. The simulation indicates that spatial sampling algorithms can

enhance the representativeness of the selected samples, and produce estimates at least as efficient as those generated by stratified sampling.

ORCID

Biagio Aragona http://orcid.org/0000-0001-8697-2932

References

Amaturo, E. and Aragona, B. (2016). La rivoluzione dei nuovi dati: quale metodo per il futuro, quale futuro per il metodo? In F. Corbisiero and E. Ruspini (Eds.), *Sociologia del Futuro*. Trento: Wolters Kluwer, 25–50.

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

Berelson, B. (1952). *Content Analysis in Communication Research*. New York: Free Press.

De Solla Price, D. (1978). Editorial statements. *Scientometrics*, *1*(1): 3–8. doi:10.1007/BF02016836

Escobar, M., Prieto, C., Barrios, D., et al. (2017). *netCoin: interactive networks with R*. https://cran.r-project.org/web/packages/netCoin/index.html

Ginsberg, J., Mohebbi, M. H., Patel, R. S., et al. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232): 1012. doi:10.1038/nature07634

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Waltham: Morgan Kaufmann.

Herring, S. C. (2009). Web content analysis: expanding the paradigm. In J. Hunsinger, L. Klastrup, and M. Allen (Eds.), *International Handbook of Internet Research*. Dordrecht: Springer, 233–249.

Latour, B. (2005). *Reassembling the Social. An Introduction to Actor-Network Theory*. Oxford: Oxford University Press.

Lazer, D., Pentland, A. S., Adamic, L., et al. (2009). Life in the network: the coming age of computational social science. *Science*, *323*(5915): 721. doi:10.1126/science.1167742

Manyika, J., Chui, M., Brown, B., et al. (2011). *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. San Francisco: McKinsey Global Institute.

Marres, N. (2017). *Digital Sociology*. Cambridge: Polity Press.

Moreno, J. L. (1934). *Who Shall Survive?* Washington: Nervous and Mental Disease Publishing Company.

Newman, M., Barabasi, A., and Watts, D. (2007). *The Structure and Dynamics of Networks*. Princeton: Princeton University Press.

Rieder, B. (2013). Studying Facebook via data extraction: the Netvizz application. *Proceedings of the 5th annual ACM web science conference* (pp. 346–355). Paris: ACM.

Rogers, R. (2013). *Digital Methods*. Cambridge: MIT Press.

Scott, J. (2012). *Social Networks Analysis*. London: Sage.

Webb, E. J., Campbell, D. T., Schwartz, R. D., et al. (1966). *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally.