Open Journal of Social Sciences, 2024, 12, 181-206
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

Big Data, Demography, and Causality

Guillaume Wunsch[1*], Catherine Gourbin[1], Federica Russo[2]

[1]Demography Centre, UCLouvain, Louvain-la-Neuve, Belgium
[2]Freudenthal Institute, Utrecht University, Utrecht, Netherlands
Email: *g.wunsch@uclouvain.be

How to cite this paper: Wunsch, G., Gourbin, C., & Russo, F. (2024). Big Data, Demography, and Causality. *Open Journal of Social Sciences, 12,* 181-206.
https://doi.org/10.4236/jss.2024.121012

Received: November 24, 2023
Accepted: January 21, 2024
Published: January 24, 2024

Abstract

The objectives of this paper are to examine to what extent Big Data are presently used in population research and to consider their potential for causal inference. After examining the characteristics and challenges of big data, the subsequent section deals with the use of big data in the study of the key demographic phenomena and is based on a literature review for the period 2015-2022 of 63 scientific journals concerned with population issues. The final section examines to what extent the use of big data could improve causal inference. Our results show that demographers continue to privilege sources of numerical data and are less prone to use digital media data or other sources such as images. Big Data can contribute to improving explanations in demography thanks to the large number of observations and variables in the data sets, especially when they can be individually linked together. Causal knowledge requires however that one can propose and test a suitable mechanism explaining why a variation in one variable produces a variation in another variable.

Keywords

Big Data, Demography, Causality, Abduction, Deduction, Induction

1. Introduction

In the past few years, several demographers have pointed out the need to consider big data in population studies. For example, Steven Ruggles (2014) has described the large amount of microdata currently available, distinguishing between “designed” data (such as censuses) and “organic” or “shallow” data, i.e. data collected for other purposes than research (*e.g.* satellite imagery, mobile phone data, …). In her presidential address to the *Southern Demographic Association*, Stephanie Bohon (2018) has argued that demographers have long collected and analysed big data but in a small way, focusing only on a subset of the data. She considers that demographers should target big *deep* data, i.e. population-generalizable data, rather than *exhaust* or *found* data created for other purposes than research. Bohon is in favour of data-driven approaches that look at the data holistically in view of detecting possible patterns. Kashyap (2021) provides an excellent review of the potentials of big data in demography, in particular the new opportunities provided by the availability of microdata from traditional sources and their linkage and by new data sources such as digital traces. The demographic literature has shown how individual-level microdata from censuses and registers expand the scope of population research and how new data sources such as traces from digital technologies can improve the understanding of human behaviors. However no study has ever thoroughly examined to what extent demographers actually use big data and for what population sub-fields. Additionally, it remains to be seen how big data can enhance the explanatory power of models in the social sciences and in population studies in particular. These issues are analyzed in this paper.

Based on the scientific papers on big data we have found using the research engine *Google Scholar*, we first examine in this paper the definitions and characteristics of big data, their sources and types. Several challenges to the use of big data are presented, and some examples given. The subsequent section, grounded on a literature review for the period 2015-2022 of 63 scientific journals concerned with population issues, deals with the use of big data in the study of the key demographic phenomena: union formation and dissolution, pregnancy and fertility, family and household, the life course, morbidity and mortality, internal (sub-national) migration and international migration. This literature review, concentrating on high-income countries, is not intended to be a complete assessment of the use of big data in population studies but a reflection of the manuscripts published in the select group of highly cited journals focusing largely on human population. The final section examines to what extent the use of big data could improve causal inference, integrating inductive, abductive, and deductive approaches.

2. Big Data: Definitions and Characteristics

2.1. Definitions

According to the literature on big data, the term ‘big data’ was first used in 1997 in the context of the storage, management and visualization of large data objects (Cox & Ellsworth, 1997). There is however no universally accepted definition of big data. Actually, many different definitions can be found in the literature. In a study of sixty-two papers that contained a definition of big data, covering the period up to September 2015, Ylijoki and Porras (2016) identified 17 different definitions of big data. For example, big data has been defined as the management and processing of any dataset that has a huge size, as large amounts of digital information, as an outcome of technology revolution where millions of people are generating huge amounts of data, as massive datasets that are generated through a variety of resources including environmental sensors, devices such as smart watches, electronic medical records, imaging, laboratory studies, and administrative data (Al-Mekhal & Khwaja, 2019).

According to Favaretto et al. (2020), the practical definitions of big data can be classified following five different dimensions, though some of these are not always present:

- Source of data: data coming from digital technologies.
- Human component: data generated from people.
- Collection: data collected for no immediate purpose (found data) or with no informed consent.
- Data processing: data needing sophisticated computational processes to be analysed.
- Problem solving tool: capable of answering questions.

Most definitions refer to at least one of the following aspects (Ward & Barker, 2013). The first is the large size of the datasets, in the number of observations *n* and often in the number of variables *p*. The second is complexity, the data being structured (potentially available in tabular form) or unstructured (such as blogs or images), often provided by various sources. The third is technologies, meaning the tools and techniques used to manage, process and analyse the datasets. Other common attributes are that *n* = all, meaning that big data relate to the entire population within the system, and that they are generated continuously.

2.2. Characteristics

Many definitions of big data refer to their characteristics and especially to the 3V, volume, velocity and variety. These attributes were first proposed by Doug Laney (2001) working at that time at the Meta Group[1]:

- *Volume* relates to the huge amount of data. It has been estimated that the digital content will reach 180 × 10^{21} bytes by the year 2025 (University of Wisconsin, 2021).
- *Velocity* refers to data creation in real-time. For example, Google receives more than 40,000 search queries every second (Ianni et al., 2021).
- *Variety* concerns the data structure (structured, semi-structured, unstructured, generated from various sources and formats, *e.g.* text, sensor data, satellite imagery, digital pictures and videos, purchase transaction records, GPS signals).

Three more V have been added since, in particular:

- *Veracity*, referring to the possible (un)reliability of the data.
- *Value*, referring to the insights and use that can be extracted from the data, the latter only becoming valuable once they are used for a specific purpose.
- *Variability*, referring to the meaning of the data being context-dependent in space and time.

Other attributes that have been described are *exhaustivity* (meaning *n* = all,

[1] Reference is also given to Gartner Inc. where Laney subsequently worked.

though "all" can be small for some populations), *relationality* (conjoining of different datasets), and *scalability* (expanding in size rapidly) (Kitchin & McArdle, 2016).

An important point made by Ylijoki and Porras (2016) is that among the characteristics of big data one should separate the data and its usage. For example, value and veracity would not be characteristics of big data *per se*, but instead reflect the usage of the data. For practicing demographers in empirical studies, veracity and value would however be among the major characteristics of the data, in the search for an explanation of the latter. This would not necessarily be the case in theoretical studies based for instance on the simulation of artificial data.

In this paper, we have considered the following characteristics as key in demography: a reliance on *found data* (wholly or partly), including digital media data, but excluding *made data* such as specific-purpose surveys, available at the observation-unit level, and concerning a very large number of observations and variables.

2.3. Data Sources and Types

Much of the literature dealing with big data refers to social media or more generally to digital media data. Though these have been used in demography (see section 4), most of the sources demographers use relate to quantitative surveys, censuses, and administrative data (Kashyap, 2021). A comparison between social surveys, administrative data and data considered as big, can be found in the literature (Kitchin & McArdle, 2016).

Data collected in quantitative surveys are *made* data in the sense that they are conducted in order to answer research questions. Though quantitative surveys can be structured and highly complex, with a large number of variables collected, they do not usually satisfy the 3V criteria concerning volume and velocity. Moreover, the sample population is known and *n* is much smaller than all but considered to be representative of the reference population.

Administrative data are *found* data, meaning that they are not originally collected for a specific research purpose but for responding to legal requirements (*e.g.* live birth certificates), for obtaining periodic information on the whole population of a country (the census), or for monitoring trends (*e.g.* natural increase from population registers). Administrative data are less complex than survey data, and are structured according to their objective. Their volume can be very large (*e.g.* individual census data for the whole population) and in some cases such as registries, data can be recorded in quasi real time. They cover their whole population of reference and classic statistical methods can be used. According to the 3V characteristics, administrative data can be assigned to the big data category (Connelly et al., 2016). Administrative registers can however lead to over-coverage if emigrations are poorly recorded; this can be a problem in migration research (Careja & Bevelander, 2018). Moreover, in all these sources some persons remain uncovered and thus undocumented. For example, population registers usually include only the legal residents of the country.

The analysis of administrative data can be considerably enhanced by linking various sources together, such as linking census data to vital registration or hospital records to taxation data. The European Nordic countries in particular have numerous registers that can be linked together. To give one example, Santavirta and Myrskylä (2015) have analysed a nationally representative sample of Finnish evacuees born in 1933-1944 by combining data collected from wartime government records with the 1950 and 1971 censuses and 1971-2011 population registers. In the European Union, datasets can often be linked using the personal identification number (PIN) given to each individual at birth or other personal identifiers such as a social security number. If personal identifiers are not available, other linkage methods (deterministic or probabilistic) can be used; the issues involved are examined in Harron et al. (2017).

Other sources of big data often concern digital media, such as digital trace data obtained from mobile phones, from social media (*e.g.* YouTube, Facebook[2], Twitter now X), online news, or from Google online searches. For example, the European Union (EU) has examined the potential of geo-referenced social-media data from Facebook and Twitter (X) to complement traditional sources for EU mobility statistics (Gendronneau et al., 2019). Acolin et al. (2022) have used consumer trace data assembled by commercial vendors[3] to produce small-area population estimates. This type of Big Data may of course come from many other sources. In the field of health care for example, clinical data are provided by electronic medical records, by diagnostic results (such as those from imagery), by molecular data, and by administrative data (such as discharge records). All these sources are *found* data, meaning that they are not collected nor designed for research purposes. They are often very large and complex, i.e. recorded in real time on various supports and with different formats. Their representativeness is usually unknown, as well as the population of reference. Classic statistical methods often cannot be applied due to the very large volume of some sources, the variety of formats (tabular, text, images, etc.) and their high velocity (such as transactions on the financial markets). The strengths and weaknesses of digital trace data are discussed in Tjaden (2021), with a focus on migration research. In sum, in the context of demography, we consider that big data are found data that are available at the observation unit level and concern a very large number of observations and variables.

3. Challenges

On the basis of the literature on big data, we first review challenges posed to big data in general, and later we provide two specific examples relevant for demography.

[2] Facebook is now part of the Meta group.

[3] Data sources are *e.g.* credit card billing statements, voter registrations, real estate tax assessments, utility records, etc.

Many problems and one solution

Using big data is not without challenges. For instance, Baro et al. (2015) point out the following ones.

- Challenges on veracity: big data often have low veracity and their representativity can be difficult to validate.
- Challenges on the workflow: in data capture, storing, cleaning, analysing, visualizing.
- Challenges in analytics: traditional methods can be ill suited for dealing with the 3V and the unstructured nature of some of the data. New computational and visualization methods are often necessary.
- Challenges on extracting meaning: many big data are not fully described nor correctly archived, and usually contain erroneous data and noise.
- Challenges on facilitating extraction access: many data do not come from open sources but from private companies, and full data access is not guaranteed. This raises the question of who get access to the data and with what constraints.

Other challenges that have to be considered include incomplete or conflicting data, data with different identifiers or formats, asynchronous data, and tampered or encrypted data (Wang, 2017). It has also been pointed out that in the field of innovative data it is the nature of the data and not their amount, the joining of different data sources, their use in a policy context, that are presently the main challenges to their use.

In the case of social media data, more problems can arise (Boyd & Crawford, 2012). For example, due to poor archiving, it can be difficult or impossible to access older data. The data cleaning procedure is not always clear. Large datasets are often unreliable and the problem is amplified when multiple datasets are used together. Furthermore, *n* is most usually not all. In other words, social media data are provided by a particular subset of all that is neither representative nor random. For instance, Facebook's CrowdTangle database only includes public Facebook pages with more than 50 K likes, public Facebook groups with 95k+ members, U.S.-based public groups with 2k+ members, and all verified profiles. It excludes among others all data or posts from private accounts. Some users create fake accounts or include fake information in their profile, for instance with respect to age or occupation. In addition, some accounts may be *bots*, i.e. softwares that automatically perform actions such as tweeting and re-tweeting. Twitter (X) estimates that around 5% of its accounts are bots (Twitter, 2021). Finally, as for all microdata, there are the issues of privacy disclosing. For example, in their use of the *WhatsApp* social network, where no public anonymized database currently exists, Rosenfeld et al. (2018) had to develop a software that encrypted the data that were taken directly from the participants' smartphones.

Some of these challenges can be countered. For example new technologies of data processing such as *Hadoop*[4] or *IBM Streams*[5] have been developed for tack-

[4] A batch-oriented processing tool allowing the distributed processing of large data sets across clusters of computers, each offering local computation and storage.

[5] A data stream management system in real time.

ling voluminous or high velocity data. New computational methods for detecting patterns and associations in the data, based on machine-learning and artificial intelligence, are now available in a data mining perspective. And confidentiality issues have lead to differential privacy outlined below. On the other hand, combining different data sources and formats remains at present a problem still to be fully resolved, and this is also the case for the low veracity and representativity challenges. In addition, the sharing of found data from private sources will require some form of legislation regulating the access to data held by the private sector.

The newer techniques of *differential privacy* may attenuate the privacy risk in big data. In 2006, Dwork et al. (2006) published a path breaking paper on privacy-preserving statistical data bases, or more generally on privacy-preserving analysis of data, where the true answer to a query is perturbed by the addition of random noise (in this case, noise coming from the Laplace distribution) and the true answer plus noise is returned to the user. They formalize in their paper the amount of noise required and propose a generalized mechanism for this purpose. The less there are people in the database, the more noise needs to be added to preserve the privacy of any individual data. There is thus a trade-off between privacy and utility. A *privacy loss parameter* denoted by ε determines how much noise is to be added: the higher the epsilon, the more accurate the data will be, but the less privacy will be preserved. For a non-technical presentation of differential privacy see Wood et al. (2020), and for detailed examples the reader is referred to e.g. U.S. Census Bureau (2021). Open-source software for statistical analysis offering the protection of differential privacy is available for instance at *OpenDP* (https://opendp.org). It can be used to build applications of privacy-preserving computations based on the 'curator model', where a trusted data collector is presumed to have access to unprotected data and wishes to protect public releases of information.

Differential privacy is quite effective against the risk of reconstruction-abetted disclosure but the methodology is not fool proof. For example, Lin and Xiao (2023) have shown that the method, in this case the TopDown Algorithm (TDA) used by the U.S. Census Bureau, does not necessarily prevent the identification of population uniques using the public census tables, such as an individual belonging to a cell in a table that has a value of one. In particular, the probability of disclosure depends upon the size of the population, its composition and the number of attributes taken into account.

Two examples from the morbidity and health field

The two examples presented here are taken from the field of morbidity and health. The first paper, by Timmins et al. (2018), reviews some contributions of found data to obesity research. The authors cover a large variety of sources that have been considered in obesity research. Retail sales have been used, for example, for monitoring nutrition intakes at the population level and for evaluating the impact of public health campaigns on food purchases. Data provided by commercial weight management programs (such as *Weight Watchers*) yield information on the effectiveness of these programs in relation to characteristics of the participants. Social media sources, such as Twitter (X), have been used *e.g.* for showing that urban areas in the U.S. with lower obesity rates more frequently discuss food, especially fruits and vegetables, and physical activities. Mobile phones and wearable technologies allow the measurement of physical activities, located in time and place thanks to GPS, in relation to characteristics of the owners. The paper also discusses the relevance of transport data, where few studies have been conducted for obesity research. The authors point out some issues: the limited data access, the validity of found data and for some sources the difficulty of linking the information to individual behaviours.

In relationship with the Covid-19 epidemic, the second paper (Yang et al., 2021) examines the so-called *Infodemics*, i.e. the diffusion of misinformation related to the disease. The authors have collected social media data from both Twitter (X) and Facebook using the same keyword list. They have then determined the credibility of the tweets and posts by "tracking the URLs linking to the domains in a pre-defined list (…) based on information provided by the *Media Bias/Fact Check* website"[6] (p. 3). Results show that low-credibility domains as a whole have a prevalence of links higher than every single high-credibility domain, though each low-credibility domain usually has a much lower prevalence than the high-credibility ones. The paper also shows, among others, that there are Infodemics super spreaders, and that there are clusters of accounts that act in a coordinated fashion to amplify Infodemics messages. One important problem pointed out by the authors is that the Twitter (X) data are based on a 10% random sample of tweets and not users, the dataset being biased towards users that are more active. The Facebook data is limited by the selection of pages and groups in CrowdTangle as indicated above.

These two papers illustrate some of the problems pointed out previously. For instance, none of the data has been initially collected for research purposes. The characteristics of the populations included in these sources can be different from those that would be observed in a random sample of the general population. Finally, access to found data is not guaranteed and, if obtainable, the information concerns only a fraction of the set of data that could be available in principle.

4. Big Data and Demography

As pointed out in the introduction, no methodical review of the actual use of big data by demographers has been published. We attempt here to fill the gap. The objectives of our literature review are twofold: 1) to identify what types of big data are used in demography and 2) to distinguish the subfields in which they are used.

Selection of journals

We first focus on the population journals that have been the most cited in the

[6] MBFC is currently the most comprehensive media bias resource on the internet.

literature, considering that these journals form the core sources of literature in demography. Most of the authors publishing in these core journals are demographers. We have used the Scopus database on the *Resurchify.com* website[7]. Using keywords "Population" and "Demography", we have first selected all journals having an *impact score*[8] greater than one. This score is a measure of the yearly average number of citations to recent articles published in a journal. This list has then been checked using the Elsevier CiteScope metric in order to see on the one hand if we had not missed some journals, and on the other hand to examine the congruence of the ranking given by both metrics. Of course, this does not mean that a less cited journal is necessarily scientifically less significant.

From this list of journals, we have kept those that we considered, based on our expertise, being focused on the key demographic processes: fertility, union formation and dissolution, mortality and morbidity, and migration. A list of 14 journals was reached on this basis. After perusal, one of them however, *Migration Studies*, fell outside the range of the article selection rules outlined below, its scope being multidisciplinary and not addressed to demography *per se*. Finally, 13 journals corresponded to all of our criteria (see Appendix 1). All the journals selected are published in English, though the authors come from a large variety of institutions worldwide. Some of the journals cover the whole field of demography, such as the journal *Demography* or the *European Journal of Population*, while others are restricted to a particular sub-field, *e.g. International Migration* or the *Journal of Population Economics.* Of course, there are some excellent papers on demography and big data in population journals that did not meet our selection criteria, and several are cited in the present article.

With increasing multidisciplinarity, demographers can publish their papers based on big data in journals other than the demographic ones. To capture these 'outer-circle' journals, we first searched on *Google Scholar* using a series of conjunctive keywords such as "big data" *and* "demography" *and* "fertility" (or mortality, or migration + mobility). This produced a huge number of sources: more or less 11,600 for fertility, 17,800 for mortality and 19,200 for migration. Moreover, after some checking, we found that many of these sources were irrelevant for the purpose of this paper. Modifying the keywords did not improve the search. We also used a database approach, with *Sociological Abstracts* for the social sciences, plus *PubMed* and *Embase* for the health field. This also led to an excessive number of papers. For example, with the filters "mortality AND register", *Embase* produced 7919 results for the U.S. and 15,556 for Europe.

As a last resort, we selected journals referenced in the bibliographies of several recent reviews dealing with data innovation and demography (Gendronneau et al., 2019; Bosco et al., 2022; Kashyap et al., 2023; Kashyap & Zagheni, 2023; Rampazzo et al., 2023) and applied to these journals the same selection criteria as those for the core demographic journals previously discussed. This added 53

[7] Websites accessed on April 20, 2022.

[8] The *impact score* is based on Scopus data and can be a little higher compared to the *impact factor* using the Web of Science data.

journals satisfying our *impact score* criterion (IS > 1) to our preliminary list of 13. Out of these journals, 50 contained at least one article corresponding to our article selection criteria (see below). These 50 journals are listed in Appendix 2.

Next to journals, demographers also publish interesting material in book chapters, proceedings, reports, and working papers. These latter sources are not considered here, as our literature review is not intended to be a complete assessment of the use of big data in demography, as pointed out in the *Introduction*, but a reflection of the manuscripts published in the rather exclusive circle of highly cited journals dealing for the most part with human population.

Selection of articles

For each of the 63 journals and for the period 2015-2022, the abstracts of all articles and the sections on *Data* and *Methods* were examined by the first two authors of this paper. In cases of uncertainty, the large majority of cases, the full contents of the papers were also examined. *Criteria for selection* were the following: a reliance on *found data* (wholly or partly), including digital media data, therefore excluding papers based on *made data* such as specific-purpose surveys (*e.g.* fertility or health surveys). Moreover, the studies could be based on census data, possibly sampled, on large-scale general-purpose surveys[9], and on population registers and administrative records, under the condition that the analyses were conducted at the observation-unit level (individual, household, …), and concerned big data volumes (arbitrarily defined here as 100,000 units at least[10]). Studies based on aggregate indicators, such as fertility rates, parity progression ratios, or life expectancies, were excluded. Once again, the focus of the papers should be on the core demographic phenomena.

Before presenting the results, some words of caution. The journals selected here are only a subset of those on the market dealing with population issues. For example, some well-known population journals have not been kept for this review because their impact score was lower than one, and they can contain papers on the topics considered here. Moreover, though we have added 50 journals to our initial list of 13, others also publish papers on the subject. Finally, the papers retained are not necessarily written by demographers, especially in the 50 "outer-circle" journals. In these, due to overlapping fields, many of the authors of the papers on mortality and morbidity, for example, come from epidemiology and the health field. To give another example, many geographers and economists are involved in migration research. Moreover, other selection criteria would lead to other choices.

The 13 *core journals*

We start with some general comments on the 13 *core demographic journals*. Few papers among those examined take a data-driven or exploratory data approach, possibly because journals prefer research founded on theoretical hypotheses. Few papers too opt for the newer methods of big data analysis based on

[9] Such as the *American Community Survey* in the U.S., replacing the decennial census long-form.

[10] Some detective work was often required for finding the number of observations in the papers. Demographers should include the value of n in all their tables of results.

machine-learning and artificial intelligence; most rely on classic econometric and statistical techniques. In these core journals, papers using information from digital technologies are rather rare, though we did find several based on mobile phone data, on Google Trends, Twitter (X), and Facebook. In many studies, the *volume* of observations reaches hundreds of thousands and even several million; demographers *do* deal with big data! On the other hand, the *velocity* and *variety* of most of the sources are low. In particular, papers in these journals rarely consider sources other than numerical data, such as images or texts, though Chakrabarti and Frye (2017) for instance have used automated text analysis for analysing the topic contents of notebooks from the *Malawi Journals Project*.

Many papers are based on linked microdata from registers and censuses, covering the whole population. This is especially true of studies on the European Nordic countries, Denmark, Norway, Sweden, and Finland. If n, the number of observations, is often very large, the same is not true of the number of variables p considered. In most studies, p is less than two dozen, with an outcome variable, some explanatory or exposure variables and a few control variables. This reflects most probably the fact that these studies are based on background knowledge and research hypotheses, and tend to privilege a parsimonious model, or that some of the exposure variables are unavailable in the datasets. Finally, demographers are not very involved in causal modelling as developed by Pearl (2000) or Heckman (2008).

Table 1 presents the cross-classification of the 192 articles that have been selected following the criteria outlined previously, according to their *sub-field* (union formation and dissolution, pregnancy and fertility, family and household, the life course, morbidity and mortality, internal (sub-national) migration, international migration[11], other) and to their *sources of data* (census only, register[12] only, linked censuses, linked registers, linked censuses and registers, digital media, other). The papers were attributed to a cell according to the source(s) of the data, on the one hand, and to the sub-field of the dependent variable(s) on the other hand, supplemented when necessary by referring to the article's keywords. The number of articles selected (192) differs from the total (214) in **Table 1** because some papers have been tagged to more than one sub-field, such as morbidity and international migration for example.

Three sub-fields stand out in the papers dealing with big data: mortality and morbidity (53 references), international migration (46), and fertility (38). Linked registers are by far the most frequent source of big data, with 89 references out of a total of 214. The next sources are Census only (47 references) followed by Register only (36). Overall, 125 analyses are based on register data and administrative records only and 20 more have recourse to linked registers and censuses. This is because many studies concern countries, such as the European Nordic ones, where many administrative registers are available and can be linked

[11] Flows and characteristics of migrants for both internal and international migration.

[12] Including administrative records.

Table 1. Summary of the selected articles according to their sub-fields and their data sources (13 core journals).

Sources Fields	Census only	Register[a] Only	Linked censuses	Linked registers	Linked censuses and registers	Digital media	Other	Total
Union (formation, dissolution)	6	7	0	12	3	0	0	28
Fertility	10	6	0	20	2	0	0	38
Family Household	6	1	1	5	1	1	0	15
Life course	4	2	1	11	1	0	0	19
Mortality Morbidity	4	14	2	23	7	2	1[b]	53
Internal migration	2	0	0	6	0	3	0	11
International migration	13	6	3	12	6	6	0	46
Other	2[c]	0	1[d]	0	0	1[e]	0	4
Total	47	36	8	89	20	13	1	214

[a]Including administrative records; [b]19 waves of NHIS linked to mortality records; [c]Ethnic diversity; effect of one-child policy; [d]Race; [e]Demography of Twitter users.

anonymously thanks to the personal identification number given to each resident.

Sub-fields differ according to the source of data. We highlight here the most salient results. Census microdata are mainly used to study international migration and fertility. Registers, either linked or not, are utilized first to study mortality and morbidity, but linked registers are also used in the second place to study fertility. Though rather rare in the core journals, studies relying on digital media are mainly concerned with mobility (international and internal migrations).

From this analysis of the core demographic journals, one has the impression that demographers use data as they have done in the past, relying mainly on censuses and registers, but taking advantage of the fact that these data are now available at the micro level and that individual linkages can be performed between sources.

The 50 *"outer-circle" journals*

These journals consider a large variety of topics, including demographic ones. The authors come from demography but also from other disciplines. Similarly to the first table, Table 2 presents a summary of the articles dealing with population topics according to their sub-fields and data sources. Once again, an article may refer to more than one sub-field.

The sub-field most concerned by far is mortality-morbidity, with 267 references out of 384, i.e. 70%. The following sub-fields are migration/mobility (internal and external) and fertility. The main sources of the data are Registers only followed by Linked registers, the latter being especially frequent in the European Nordic countries. Digital media are the third most common source of data in these outer-circle articles. Registers (all categories) are mostly exploited for the

Table 2. Summary of the selected articles according to their sub-fields and their data sources (50 outer-circle journals).

Sources Fields	Census only[a]	Register[b] only	Linked censuses	Linked registers	Linked censuses and registers	Digital media	Other[c]	Total
Union (formation, dissolution)	2	1		4	1	1		9
Fertility	1	10		7	1	3		22
Family Household		1		1				2
Life course	1			1				2
Mortality Morbidity	6	93		73	32	40	23	267
Internal migration	4	2	4	4	2	24		40
International migration	5	11	1	7	6	7	4	41
Other[d]						1		1
Total	19	118	5	97	42	76	27	384

[a]Including micro-census; [b]Including administrative records and follow-up cohort studies; [c]Pooled surveys, or linked survey(s) and registers; [d]Race.

study of mortality-morbidity, while Digital media are utilized for the latter[13] but also for the study of migration and mobility.

Comparing the two tables, some major differences stand out. For the 13 core demographic journals, the category Census only is much more important than in Table 2, while the opposite is true for Digital media. Moreover, the latter are mainly employed in the core journals for migration studies while in the outer-circle journals digital media are used too for studying morbidity. This shows that studies based on digital media are more often published in journals with a larger scope than the core demographic ones. Finally, the 13 core journals deal more with union formation and dissolution, and with fertility, than the 50 outer-circle journals.

5. Big Data and Causality

Causal questions are at the core of demographic research. However, the way in which demographers have explicitly engaged with causal questions and used causal models has had ups and downs in the past decades. Big data complicate the picture further, because they are often considered not suitable for establishing causal relations but only correlations. Whence our question: can causal research profit from the availability of big data? Searching for cause-effect relations is not only important for scientific reasons, a good explanation usually requiring the knowledge of the causes of an effect or the effects of a cause. It is also crucial for policy-makers, as the recent COVID epidemic has shown. What are the origins of this respiratory disease? What would be the consequences of a mandatory vaccination program? Policy-makers cannot discard the causation issue, as informed decisions require being based on evidence and knowing the probable consequences of an intervention.

[13] Many are concerned with the Covid-19 epidemic.

During the past few years, an empiricist perspective on the nature of science has been developed in response to the advent of Big Data analytics, leading to a possible paradigm shift. See for example the discussion in Symons and Alvarado (2016). According to the empiricist camp, Big Data analytics would lead to the "death of theory", as its capacity to detect patterns and associations in the data, and to build predictive models, could replace hypotheses born from *a priori* theory. The hypothetico-deductive method common to much of the scientific enterprise would be replaced by a purely data-driven inductive approach where hypotheses and explanations are generated *ex post* from the data themselves. As Kitchin (2014) has however observed, "it is one thing to identify patterns; it is another thing to explain them". In other words, data do not speak for themselves and have to be interpreted. In addition, even if the data set is big, nothing guarantees that it contains all the variables that have to be controlled for to avoid confounding.

Kitchin's remark contrasts with the provocative statement made by Anderson (2008) a few years earlier, predicting that Big Data would mark 'the end of theory' and the beginning of an era in which (Big) data would speak for themselves. Regardless of the prophecy of Anderson, the question of how to get 'causal facts' from data is a long-standing debate in the philosophy of causality and in methodological debates across the sciences (Illari & Russo, 2014). What is at stake is the ability and possibility of disentangling claims about correlation and about causation, and thus to make the step from description to explanation of a given phenomenon.

The question of providing an epistemology for Big Data practices, able to account for the step from correlation to causation, is precisely the task taken up by Pietsch (2021). His point is that, although the arguments of Anderson are clearly flawed, we should still strive to provide an elaborated argument for why they are flawed, which is the main objective of his book. His reconstruction of Big Data practices sees in inductivism, and specifically in *variational induction*, the key inferential tool used to infer causes from correlations in very large data sets. Variational induction, in Pietch's view, is part of the epistemological and methodological tradition initiated by John Stuart Mill, and the reliance on the notion of variation resonates with, and builds upon, the variational epistemology for causal modelling in the social sciences as is developed by Russo (2009).

It is worth clarifying that Pietsch begins with discussing 'induction' in the classical sense, but ultimately his aim is to provide an understanding of inductive practices in the contemporary methodological context, and so his view on 'variational induction' does not squarely coincide with the Baconian view of induction (see *e.g.* Courgeau et al., 2014). The main point of Pietsch is that it is incorrect to claim that Big Data abandons causal inference altogether. Instead, he shows that variational induction is precisely about causal relationships, and it is this inferential tool that is at work in Big Data. Induction is often associated with "eliminativism", namely we proceed by eliminating all but the true causes, and whatever remains is, by induction, the sought cause. However, Pietsch shows that eliminative induction is a kind of misnomer, and we should instead focus on the method of difference and of agreement: it is by *comparing circumstances* that we can proceed with further selecting or eliminating putative causes. In this stage of comparison, *variation* has a more fundamental role than elimination. In the analysis of Pietsch, the vast majority of machine learning algorithms, designed to analyse big data sets, rely precisely on this rationale of variation, for instance to determine the relevance or irrelevance of predictor variables. It is also interesting to note that, in Pietch's view, the highly predictive success of Big Data practices also relies on causal relationships, albeit "approximate" or probabilistic in character, and this goes counter the other widespread view that causation ceases to have a central role in the era of Big Data.

Next to induction, philosophers of science are also reviving a discourse on the role of abduction in a number of scientific contexts (Aliseda, 2006), and especially for its potential application to medical reasoning, such as in diagnosis or in public health emergency contexts (Barés Gómez & Fontaine, 2021, 2022). Abduction has several meanings, such as taking someone away by force. In the present essay, we refer of course to its meaning in the philosophical literature. The very definition of abduction is, in logic and philosophical logic, not unanimously agreed upon. Simply put, like induction, abductive inference is ampliative, in the sense that the conclusion contains *new* information with respect to the premises, and for this reason it is not certain, but probable inference (unlike deduction). One characteristic aspect of abductive inferences, unlike inductive ones, is to generate hypotheses to be further considered for empirical evaluation. Specifically, the approach to abduction that is adopted is that of Gabbay and Woods (2005) and of Magnani (2017) that preserve the tripartite distinction between deduction, induction, and abduction, and defend the specificity of abduction for the *generation* of hypotheses.

This attention to the abductive processes is happening in demography too. Demographers have recently suggested adopting abductive reasoning in their field, either instead of or in addition to the usual deductive or inductive approaches (Hauer & Bohon, 2020; Bijak, 2022). Contrary to the empiricist perspective, the abductive perspective does rely on prior knowledge. It seeks to propose the most plausible explanation for a novel event or pattern observed. More specifically, if C is observed, abduction consists in selecting a hypothesis A from one's background knowledge or theory, deemed the most plausible for the case at hand, such that if A is true then C is explained (Catellin, 2004). However, there can be other and better causes of C than A. Abduction thus also requires testing the validity of the proposed explanation A and comparing A to other possible causes. Abduction therefore links inductive and deductive approaches. Doctors use it in practice when proposing an explanation of the symptoms observed in a patient, based on their knowledge of the causal relations between diseases and symptoms, and more generally by scientists when invoking an explanation for novel patterns discovered in an exploratory analysis of the data.

The inferential scheme of abduction seems very simple and straightforward.

In practice, however, abductive practices are very complex because it is far from obvious to choose the relevant background knowledge. Moreover, any relevant background knowledge may lead to a number of conclusions, and the justification of one over another one is precisely what makes abduction fundamentally uncertain and hypothetical. It is worth noting that variational induction or abductive reasoning never work "on their own", but are always part of a broader inferential framework at work in techno-scientific contexts: *hypothetico-deductivism*. It is an illusion to think that we can perform fully "agnostic" searches for correlations in big data sets, and likewise it is an illusion that we "simply" deduce (in a strict sense) conclusions from general statements. Scientists proceed via hypothetico-deductive approaches, in the sense that the state of hypothesis formulation always leads to subsequent steps such as data collection and analysis. At the same time, the meaning of "deduction" here does not coincide with logical deduction, but more loosely refers to any inferential step that takes scientists from hypotheses to some conclusion, and so it may include variational induction as well as abductive practices. The rise of algorithmic approaches for data analysis makes more pressing than ever to reconsider broad inferential practices at work in empirical research. Even if we can automatize part of the process of data analysis, we need to remember that behind algorithms there are always scientists (human beings in flesh and bones) that perform such inferences.

Though computers can easily detect patterns and associations, they are not (yet?) capable of proposing novel explanations for the correlations observed. This is however a strenuous attempt from scientists and especially machine learning developers to code an algorithm that is able to come up with theories and explanations, just as *human* scientists do. A pretty successful attempt in this respect was "*Eureqa*", that apparently was able to reproduce the laws of physics through evolutionary search based on AI, in a format quite close to Newton's formulation, within hours of being fed with data (Schmidt & Lipson, 2009; *Eureqa*, 2022). It remains of course to be seen the extent to which this algorithm is really working in a fully agnostic way, and instead how much background knowledge—of any kind—does play a role in the whole process. Differently put, attempts such as *Eureqa* still neglect that *humans* have done the coding, and so computers do not discover or explain anything *just* on their own.

In our previous diagnosis example, a computer could use the huge information on background knowledge and theory possibly contained in its memory or cloud storage to inform a doctor about the most common diseases associated with the symptoms observed and propose the most probable solutions. There is presently an important line of research on this topic. In the case of such computer-assisted research, computers would not only detect correlations but also suggest plausible mechanisms linking effects to their possible causes, based on background knowledge. These proposals would then have to be evaluated and tested in practice. As the proposals rely on background knowledge, i.e. on past results, novel explanations cannot however be derived by this approach.

These remarks about the ability and possibilities that computers come up with explanations on their own are part of a larger discourse on the *partnership* between humans and machines in the process of producing knowledge. As Russo (2022) explains, it is wrong to consider that humans, as epistemic agents, produce knowledge (analyse data, explain, predict, …) alone *and* that we can simply read off results from the analyses performed by machines or let them take care of the whole scientific process. It is instead crucial to realize how much scientists, as *human* epistemic agents, act *together* with technical artefacts, whether these are simple square and compasses or sophisticated algorithms for the analysis of big data sets. This is all to say that, without buying into the hype of Big Data or of the full automation of the scientific process, it remains true that Big Data (and the whole computational and algorithmic infrastructure that goes with it) are an opportunity for enlarging the scope of what we can know about social phenomena with pre-Big Data methods and techniques.

6. Discussion

Big Data and demography

Our literature review has shown that, in the field of Big Data, demographers analyse huge amounts of microdata coming from censuses (or equivalents) and from various administrative registers. In many articles, the *volume* of observations reaches hundreds of thousands and even several million. On the other hand, the *velocity* and *variety* of the sources are low. During the past years, individual-level anonymized data from censuses and registers have become increasingly available and demographers have taken advantage of this situation. In addition, more and more national institutes are now linking data sources together, namely census with registers, census with census, and registers with registers. Countries where a personal identification number is attributed to each resident are in a more favourable position for this purpose. The advantage of these official sources is that they are usually freely available, cover the entire population, and provide accurate information. Though these sources have not been set up for research purposes, the amount of information they can yield is enormous, especially in the case of record-linkage. As population can change fast or slow, Billari (2022) has made a case for a demographic data collection that would take the speed of population change into account, in particular migration inflows and outflows at various levels of spatial scale.

As **Table 1** shows, research on mortality and morbidity, for instance, has greatly benefited from record linkage. Fertility studies have also taken advantage of the situation. In these 13 core demographic journals, research has been less concerned with digital media sources such as Facebook or X. The most interesting contribution is the use of trace data from mobile phones for estimating international or internal migration, and refugee mobility. A downside is that migrants may change phones or SIM cards after migrating to another country, and locations are only recorded when calls are made (Tjaden, 2021). The less frequent use of digital media sources compared to traditional ones can be due to the fact that most of the social media sources, such as Facebook or X, do not yield 'hard' demographic facts but focus instead on interactions and communication between persons, on socializing with friends, on personal interests and sentiments, and on shared comments. In addition, digital media usually contain very few individual characteristics that can be used for demographic purposes. Linkage with other sources is in the most cases impossible. Contrary to the core demographic journals, in the outer-circle journals of **Table 2** digital media are an important source of data, not only for studying migration and mobility but especially for analysing morbidity and health. This seems to show that population scientists tend to publish their studies based on digital media in journals with a larger scope than the core demographic ones.

Our results have shown that demographers are not very involved in causal modelling. This could be due to the fact that research in demography still has often recourse to single equation models where an outcome variable is related to a set of explanatory and control variables through some functional form of relation. A single equation model cannot spell out the structure of relationships among the variables. This requires the use of multiple equation models that can deal with more complex designs (Wunsch & Gourbin, 2020). This is probably even more crucial when using large data sets with many variables that have to be structured to give causal meaning.

Big Data and causality

What can big data tell us about causality? Two cases can be distinguished, according to the fact that one is following either a *hypothetico-deductive* or a *data-driven* approach. In the first case, most attempts at causal modelling in the social sciences now refer to Judea Pearl's methodology based on *directed acyclic graphs* or DAGs (Pearl, 2000). A DAG is a subset (usually small) of the p(p-1)/2 edges[14] of a complete undirected graph taking all the *p* variables of the dataset into account, where only the postulated directed links are maintained according to one's theory and background knowledge. A DAG graphically represents the recursive decomposition of a multivariate distribution. DAGs are used to represent causal relationships and are also known as *Bayesian networks* because of the subjective nature of the input information, the reliance on Bayes' conditioning for updating information, and the distinction between causal and evidential modes of reasoning (Pearl, 2000: p. 14). An example in the field of reproductive health is given in Gourbin et al. (2017). In this model, a DAG expresses the links postulated among variables, where each variable or node in the graph depends upon the variables "upstream" in the graph, in the absence of feedback effects. Each directed edge (arrow) or link represents a putative causal effect and each endogenous variable is conditioned on its immediate causes, i.e. the variables that have a direct effect upon this endogenous variable.

For a general causal framework relying on DAGs, see Russo et al. (2019). This framework endorses the hypothetico-deductive method and interprets the re-

[14] Or twice that number if a bi-directed complete multigraph is considered.

cursive decomposition of the multivariate distribution as a mechanism composed of various sub-mechanisms. In current practice however, as exemplified by many papers consulted for this article, a more traditional approach is usually adopted by regressing an outcome variable on a number of postulated explanatory or exposure variables and on a series of controls, without specifying the causal order or structure among the variables.

Big data can help improving the hypothetico-deductive methodology in several ways. On the one hand, a large number *n* of observations increases the precision of the estimates and the power of hypothesis tests. On the other hand, if *n* is very large, even small differences of no theoretical interest will be "statistically significant", blurring the causal picture. Statistical significance should not be confused with theoretical significance (Bijak, 2019). Following Titiunik (2015), a large number of observations can also allow for a wider range of estimation methods that would be unreliable with fewer observations. A large *n* also enables studying small subpopulations that would be overlooked in *e.g.* sample surveys. For example, the Indonesian 2010 census provides information on 964 ethnic groups, some being very small (Guilmoto, 2015).

Though much of the focus is usually on the number of observations, a large number *p* of variables helps in better describing the event being studied and in reducing omitted-variable bias. However, as the number of variables becomes very high, the differences among individuals will increase, as their profiles will differ more and more. As each individual becomes increasingly unique, will grouping *e.g.* life courses together using sequence analysis still have causal meaning? Theoretical reflection on the plausible subset of causal variables will be especially required in these circumstances.

As pointed out above, more and more individual linkages are now performed among different data sources covering the whole population (censuses, registers, administrative records, …) in longitudinal or intergenerational studies. In addition, recourse to machine learning and artificial intelligence should be expanded, as they are probably better suited for exploring large n x p data matrices than more traditional techniques. For example, Aizawa (2020) has used machine learning—in this case the *random forest* algorithm—to estimate the relationship between early-life circumstances and child height among children in Vietnam, Peru, Ethiopia, and India. Arpino et al. (2022) have used random survival forests to analyse data on married or cohabiting couples who participated in the German Socio-Economic Panel Survey, in order to find predictors of union dissolution and to explore nonlinearities and nonadditivities in the links between these predictors and union dissolution. This decision-tree approach is especially robust to overfitting, multicollinearity and statistical noise (Breiman, 2001).

Big data can also help by providing new sources of information, such as digital trace data from mobile phones that can be used for estimating migration and mobility, smart watches for monitoring physical activity, Facebook and Google for measuring for instance intentions and interests, or sources for obtaining contextual characteristics of the population such as satellite imagery. For example, Billari et al. (2013) have examined fertility-related searches performed through the Google web search engine using three keywords: "maternity", "ovulation" and "pregnancy", as indicators of fertility-related interests of web users. An issue here is that the keywords one uses to capture intentions and interests may vary over time and region.

In a *data-driven* research, the purpose can be either an exploratory analysis and description of the data or an attempt to derive some causal meaning from the data themselves. In both cases, one must consider that a p-variable undirected graph may be expressed by a very high number of forms or configurations of the associations among the variables that could possibly represent interesting patterns of interrelations. For example, with six variables or nodes, there are 156 possible forms. With seven variables, the number increases to 1044, and with 10 nodes, one exceeds 12 million possible configurations (Curien, 2022). This shows that dimension-reduction techniques should be applied even when one is considering a small number of variables.

7. Conclusion

To conclude, can data speak for themselves? One can presume that among the various subgraphs that will be highlighted by the associations among nodes, only a few will possibly make causal sense. The point is similar to identifying communities in complex networks (see Xiang et al., 2009). An association may result from causation, from observation bias, confounding, or chance. The issue is not merely that of controlling for confounders in the dataset, i.e. examining conditional associations rather than marginal ones, in order to exclude spurious correlations. In a data-driven approach, the problem consists in proposing and testing a suitable causal mechanism that can explain *why* a wiggle observed in one variable produces a wiggle in another variable, observation bias and confounding being under control. This still requires human intervention and a research design based on induction, abduction, and deduction.

Acknowledgments

Comments and suggestions from an anonymous reviewer are gratefully acknowledged.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Acolin, A., Decter-Frain, A., & Hall, M. (2022). Small-Area Estimates from Consumer Trace Data. *Demographic Research, 47,* 843-882.
https://doi.org/10.4054/DemRes.2022.47.27

Aizawa, T. (2020). Trajectory of Inequality of Opportunity in Child Height Growth: Evidence from the Young Lives Study. *Demographic Research, 42,* 165-202. https://doi.org/10.4054/DemRes.2020.42.7

Aliseda, A. (2006). *Abductive Reasoning. Logical Investigations into Discovery and Explanation*. Springer. https://doi.org/10.1007/1-4020-3907-7

Al-Mekhal, M., & Khwaja, A. A. (2019). A Synthesis of Big Data Definitions and Characteristics. In 2*019 International Conference on Computational Sciences and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)* (pp. 314-322). IEEE. https://doi.org/10.1109/CSE/EUC.2019.00067

Anderson, C. (2008). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. Wired. https://www.wired.com/2008/06/pb-theory/

Arpino, B., Le Moglie, M., & Mencarini, L. (2022). What Tears Couples Apart: A Machine Learning Analysis of Union Dissolution in Germany. *Demography, 59,* 161-186. https://doi.org/10.1215/00703370-9648346

Barés Gómez, C., & Fontaine, M. (2021). Medical Reasoning in Public Health Emergencies. Below High Standards of Accuracy. *Teorema: Revista Internacional de Filosofía, 40,* 151-173.

Barés Gómez, C., & Fontaine, M. (2022). Medical Reasoning and the GW Model of Abduction. In L. Magnani (Ed.), *Handbook of Abductive Cognition* (pp. 1-26). Springer. https://doi.org/10.1007/978-3-030-68436-5_14-1

Baro, E., Degoul, S., Beuscart, R., & Chazard, E. (2015). Toward a Literature-Driven Definition of Big Data in Healthcare. *BioMed Research International, 2015,* Article ID: 639021. https://doi.org/10.1155/2015/639021

Bijak, J. (2019). Editorial: P-Values, Theory, Replicability, and Rigour. *Demographic Research, 41,* 949-952. https://doi.org/10.4054/DemRes.2019.41.32

Bijak, J. (2022). *Towards Bayesian Model-Based Demography. Agency, Complexity and Uncertainty in Migration Studies*. Springer. https://doi.org/10.1007/978-3-030-83039-7

Billari, F. C. (2022). Demography: Fast and Slow. *Population and Development Review, 48,* 9-30. https://doi.org/10.1111/padr.12464

Billari, F. C., D'Amuri, F., & Marcucci, J. (2013). Forecasting Births Using Google. In PAA, *The Population Association of America Annual Meeting, Session 155: Methods and Models in Fertility Research* (pp. 1-30). Population Association of America.

Bohon, S. A. (2018). Demography in the Big Data Revolution: Changing the Culture to Forge New Frontiers. *Population Research and Policy Review, 37,* 323-341. https://doi.org/10.1007/s11113-018-9464-6

Bosco, C., Grubanov-Boskovic, S., Iacus, S.M., Minora, U., Sermi, F., & Spyratos, S. (2022). *Data Innovation in Demogaphy, Migration and Human Mobility*. EUR30907 EN, Publications Office of the European Union.

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society, 15,* 662-679. https://doi.org/10.1080/1369118X.2012.678878

Breiman, L. (2001). Random Forests. *Machine Learning, 45,* 5-32. https://doi.org/10.1023/A:1010933404324

Careja, R., & Bevelander, P. (2018). Using Population Registers for Migration and Integration Research: Examples from Denmark and Sweden. *Comparative Migration Studies, 6,* Article No. 19. https://doi.org/10.1186/s40878-018-0076-4

Catellin, S. (2004). L'abduction: Une pratique de la découverte scientifique et littéraire. *Hermès, La Revue, 39,* 179-185. https://doi.org/10.4267/2042/9480

Chakrabarti, P., & Frye, M. (2017). A Mixed-Methods Framework for Analyzing Text Data: Integrating Computational Techniques with Qualitative Methods in Demography. *Demographic Research, 37,* 1351-1382. https://doi.org/10.4054/DemRes.2017.37.42

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The Role of Administrative Data in the Big Data Revolution in Social Science Research. *Social Science Research, 59,* 1-12. https://doi.org/10.1016/j.ssresearch.2016.04.015

Courgeau, D., Bijak, J., Franck, R., & Silverman, E. (2014). Are the Four Baconian Idols Still Alive in Demography? *Revue Quetelet/Quetelet Journal, 2,* 31-59. https://doi.org/10.14428/rqj2014.02.02.02

Cox, M., & Ellsworth, D. (1997). Managing Big Data for Scientific Visualization. In IEEE, *Proceedings of the 8th Conference on Visualization'97* (pp. 5-17). IEEE Computer Society Press.

Curien, N. (2022). Transitions de phase dans les graphes aléatoires: Une preuve inespérée. *La Recherche, 570,* 107-111.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In S. Halevi, & T. Rabin (Eds.), *Theory of Cryptography. TCC 2006. Lecture Notes in Computer Science* (Vol. 3876, pp. 265-284). Springer. https://doi.org/10.1007/11681878_14

Eureqa (2022). *Reverse Engineering Dynamical Systems.* https://www.creativemachineslab.com/eureqa.html

Favaretto, M., De Clercq, E., Schneble, C. O., & Elger, B. S. (2020). What Is Your Definition of Big Data? Researchers' Understanding of the Phenomenon of the Decade. *PLOS ONE, 15,* e0228987. https://doi.org/10.1371/journal.pone.0228987

Gabbay, D., & Woods, J. (2005). *The Reach of Abduction Insight and Trials* (Vol. 2). Elsevier Science.

Gendronneau, C., Wiśniowski, A., Yildiz, D., Zagheni, E., Fiorio, L., Hsiao, Y., Stepanek, M., Weber, I., Abel, G., & Hoorens, S. (2019). *Measuring Labour Mobility and Migration Using Big Data. Exploring the Potential of Social-Media Data for Measuring EU Mobility Flows and Stocks of EU Movers.* European Commission.

Gourbin, C., Wunsch, G., Moreau, L., Guillaume, A., & ECAF Team (2017). Direct and Indirect Paths Leading to Contraceptive Use in Urban Africa. *Revue Quetelet/Quetelet Journal, 5,* 33-71. https://doi.org/10.14428/rqj2017.05.01.02

Guilmoto, C. Z. (2015). Mapping the Diversity of Gender Preferences and Sex Imbalances in Indonesia in 2010. *Population Studies, 69,* 299-315. https://doi.org/10.1080/00324728.2015.1091603

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., & Goldstein, H. (2017). Challenges in Administrative Data Linkage for Research. *Big Data & Society, 4,* 1-12. https://doi.org/10.1177/2053951717745678

Hauer, M. E., & Bohon, S. A. (2020). Causal Inference in Population Trends: Searching for Demographic Anomalies in Big Data. https://doi.org/10.31235/osf.io/xn2v9

Heckman, J. J. (2008). Econometric Causality. *International Statistical Review, 76,* 1-27. https://doi.org/10.1111/j.1751-5823.2007.00024.x

Ianni, M., Masciari, E., & Sperlí, G. (2021). A Survey of Big Data Dimensions vs Social Networks Analysis. *Journal of Intelligent Information Systems, 57,* 73-100. https://doi.org/10.1007/s10844-020-00629-2

Illari, P., & Russo, F. (2014). *Causality: Philosophical Theory Meets Scientific Practice.* Oxford University Press.

Kashyap, R. (2021). Has Demography Witnessed a Data Revolution? Promises and Pitfalls of a Changing Data Ecosystem. *Population Studies, 75,* 47-75. https://doi.org/10.1080/00324728.2021.1969031

Kashyap, R., & Zagheni, E. (2023). Chap. 17. Leveraging Digital and Computational Demography for Policy Insights. In E. Bertoni, M. Fontana, L. Gabrielli, S. Signorelli, & M. Vespe (Eds.), *Handbook of Computational Social Science for Policy* (pp. 327-343). Springer. https://doi.org/10.1007/978-3-031-16624-2_17

Kashyap, R., Rinderknecht, R. G., Akbaritabar, A., Alburez-Guterriez, D., Gil-Clavel S., Grow, A., Kim, J., Leasure, D. R., Lohmann, S., Negraia D. V., Perrotta, D., Rampazzo, F., Tsai, C.-J., Verhagen, M. D., Zagheni, E., & Zhao, X. (2023). Digital and Computational Demography. In J. Skopek (Ed.), *Research Handbook on Digital Sociology* (pp. 47-85). Edward Elgar Publishing. https://doi.org/10.4337/9781789906769.00010

Kitchin, R. (2014). Big Data, New Epistemologies and Paradigm Shifts. *Big Data & Society, 1,* 1-12. https://doi.org/10.1177/2053951714528481

Kitchin, R., & McArdle, G. (2016). What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets. *Big Data & Society, 3,* 1-10. https://doi.org/10.1177/2053951716631130

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. *Application Delivery Strategies,* Meta Group, 3 p.

Lin, Y., & Xiao, N. (2023). Assessing the Impact of Differential Privacy on Population Uniques in Geographically Aggregated Data: The Case of the 2020 U.S. Census. *Population Research and Policy Review, 42,* Article No. 81. https://doi.org/10.1007/s11113-023-09829-4

Magnani, L. (2017). *The Abductive Structure of Scientific Creativity.* Springer. https://doi.org/10.1007/978-3-319-59256-5

Pearl, J. (2000). *Causality. Models, Reasoning, and Inference.* Cambridge University Press.

Pietsch, W. (2021). Big Data. In *Elements Philosophy of Science.* Cambridge University Press. https://doi.org/10.1017/9781108588676

Rampazzo, F., Rango, M., & Weber, I. (2023). Chap. 18. New Migration Data: Challenges and Opportunities. In E. Bertoni, M. Fontana, L. Gabrielli, S. Signorelli, & M. Vespe (Eds.), *Handbook of Computational Social Science for Policy* (pp. 345-359). Springer. https://doi.org/10.1007/978-3-031-16624-2_18

Rosenfeld, A., Sina, S., Sarne, D., Avidov, O., & Kraus, S. (2018). WhatsApp Usage Patterns and Prediction of Demographic Characteristics without Access to Message Content. *Demographic Research, 39,* 647-670. https://doi.org/10.4054/DemRes.2018.39.22

Ruggles, S. (2014). Big Microdata for Population Research. *Demography, 51,* 287-297. https://doi.org/10.1007/s13524-013-0240-2

Russo, F. (2009). *Causality and Causal Modelling in the Social Sciences.* Springer. https://doi.org/10.1007/978-1-4020-8817-9

Russo, F. (2022). *Techno-Scientific Practices: An Informational Approach.* Rowman & Littlefield Publishers.

Russo, F., Wunsch, G., & Mouchart, M. (2019). Causality in the Social Sciences: A Structural Modelling Framework. *Quality & Quantity, 53,* 2575-2588. https://doi.org/10.1007/s11135-019-00872-y

Santavirta, T., & Myrskylä, M. (2015). Reproductive Behavior Following Evacuation to

Foster Care during World War II. *Demographic Research, 33,* 1-30. https://doi.org/10.4054/DemRes.2015.33.1

Schmidt, M., & Lipson, H. (2009). Distilling Free-Form Natural Laws from Experimental Data. *Science, 324,* 81-85. https://doi.org/10.1126/science.1165893

Symons, J., & Alvarado, R. (2016). Can We Trust Big Data? Applying Philosophy of Science to Software. *Big Data & Society, 3,* 1-17. https://doi.org/10.1177/2053951716664747

Timmins, K. A., Green, M. A., Radley, D., Morris, M. A., & Pearce, J. (2018). How Has Big Data Contributed to Obesity Research? A Review of the Literature. *International Journal of Obesity, 42,* 1951-1962. https://doi.org/10.1038/s41366-018-0153-7

Titiunik, R. (2015). Can Big Data Solve the Fundamental Problem of Causal Inference? *Political Science & Politics, 48,* 75-79. https://doi.org/10.1017/S1049096514001772

Tjaden, J. (2021). Measuring Migration 2.0: A Review of Digital Data Sources. *Comparative Migration Studies, 9,* Article No. 59. https://doi.org/10.1186/s40878-021-00273-x

Twitter (2021). Four Truths about Bots. *Common Thread.*

U.S. Census Bureau (2021). *Disclosure Avoidance for the* 2020 *Census: An Introduction.* U.S. Government Publishing Office.

University of Wisconsin (2021). *What Is Big Data?* Data Sciences. https://datasciencedegree.wisconsin.edu/data-science/what-is-big-data/

Wang, L. (2017). Heterogeneous Data and Big Data Analytics. *Automatic Control and Information Sciences, 3,* 8-15. https://doi.org/10.12691/acis-3-1-3

Ward, J. S., & Barker, A. (2013). Undefined by Data: A Survey of Big Data Definitions. arXiv:1309.5821v1

Wood, A., Altman, M., Nissim, K., & Vadhan, S. (2020). Designing Access with Differential Privacy. In S. Cole, I. Dhaliwal, A. Sautmann, & L. Vilhuber (Eds.), *Handbook on Using Administrative Data for Research and Evidence-Based Policy.* https://cyber.harvard.edu/story/2021-02/designing-access-differential-privacy

Wunsch, G., & Gourbin, C. (2020). Causal Assessment in Demographic Research. *Genus, 76,* Article No. 18. https://doi.org/10.1186/s41118-020-00090-7

Xiang, B., Chen, E.-H., & Zhou, T. (2009). Finding Community Structure Based on Subgraph Similarity. In S. Fortunato, G. Mangioni, R. Menezes, & V. Nicosia (Eds.), *Complex Networks. Studies in Computational Intelligence* (pp. 73-81). Springer. https://doi.org/10.1007/978-3-642-01206-8_7

Yang, K. C., Pierri, F., Hui, P. M., Axelrod, D., Torres-Lugo, C., Bryden, J., & Menczer, F. (2021). The COVID-19 Infodemic: Twitter versus Facebook. *Big Data & Society, 8.* https://doi.org/10.1177/20539517211013861

Ylijoki, O., & Porras, J. (2016). Perspectives to Definition of Big Data: A Mapping Study and Discussion. *Journal of Innovation Management, 4,* 69-91. https://doi.org/10.24840/2183-0606_004.001_0006

Appendix 1. List of Core Demographic Journals (By Alphabetical Order) and Number of Selected Papers

• Asian Population Studies	4
• Comparative Migration Studies	1
• Demographic Research	44
• Demography	23
• European Journal of Population	24
• Genus	3
• International Migration	8
• International Migration Review	10
• Journal of Population Economics	33
• Population and Development Review	9
• Population Research and Policy Review	19
• Population Studies	13
• Studies in Family Planning	1

	192

Appendix 2. List of 'Outer-Circle' Journals (By Alphabetical Order) and Number of Selected Papers

American Behavioral Scientist	2	Journal of Urban Technology	4
American Journal of Sociology	4	Journal of Medical Internet Research	10
BMC Public Health	9	JMIR Public Health Surveillance	20
Cartography and Geographic Information Science	2	Journal of the Royal Society Interface	1
Comparative Migration Studies	2	Machine Learning and Knowledge Extraction	1
Environmental Pollution	9	Migration Letters	6
Environmental Research	10	Nature Human Behaviour	3
Environmental Research Letters	3	Neural Computing and Applications	1
EPJ Data Science	6	NPJ Digital Medicine	5
European Sociological Review	3	Palgrave Communications	2
Frontiers in Public Health	5	Personality and Individual Differences	1
Geospatial Health	3	Pervasive and Mobile Computing	2
Health Economics	13	PloS One	63
Health, Education & Behavior	3	Population, Place and Space	20
IEEE Pervasive Computing	1	Quality & Quantity	3
International Journal of Data Sciences and Analytics	4	Remote Sensing	1
International Journal of Geographical Information Science	1	Scientific Reports	13
International Journal of Health Geographics	6	Sensors	1
ISPRS International Journal of Geo-information	5	Social Science Computer Review	1
International Journal of Manpower	2	Social Science Research	6
International Journal of Public Health	15	Sociological Research Online	1
Journal of Development Economics	2	SSM Public Health	62
Journal of Geographical Systems	1	Sustainability	2
Journal of Human Behavior in the Social Environment	1	The Professional Geographer	3
Journal of Maps	8	Transportation	5
		Total	357