Journal of Population Research (2018) 35:319–324
https://doi.org/10.1007/s12546-018-9214-y

EDITORIAL

Migration and immigration: recent advances using linked administrative data

Michael Haan[1] **· James Ted McDonald**[2]

Published online: 30 November 2018

Introduction

We have long used data to understand and govern ourselves. Almost 1000 years ago, King William I ordered that a survey be taken of England after he conquered it, the results of which became known as the *Domesday Book.* The Book chronicled current and past owners of England's land holdings, the land's value, tax assessments, and the number of peasants, ploughs, livestock, and other resources on the property at the time. Location was also recorded, yielding geo-coded information on 13,418 locations across the country.

Domesday was significant for several reasons. First, it was one of the largest, if not the largest, data collection exercises conducted up to that point. William's exact motivations for such an undertaking are unclear, but it is likely that he was hoping to count the number of soldiers, assess the feudal obligations of peasants, and, almost certainly, increase taxes to pay for the building boom that he was fuelling. Nothing was missed—hence its name 'Domesday', or day of judgement, for all English persons.

Second, *Domesday* was also significant as an administrative exercise. The survey was largely conducted in 1086 and must have required an army of information gatherers, trained for the task, scouring the countryside. William's 'data army' would likely have rivalled most battalions of the day in size, yet would not have the same sort of organisational legacy. Planning *Domesday* would have been quite an organisational feat.

Third, the data were gathered before many of the techniques that are well-suited for its analysis. *Domesday* provides basic counts of holdings, cursory valuations of property and chattel, and some other general overview comments, all of which seem to stem from casual observation than any systematic analysis.

✉ Michael Haan
mhaan2@uwo.ca

[1] The University of Western Ontario, London, ON, Canada

[2] The University of New Brunswick, Fredericton, NB, Canada

To summarise, the *Domesday Book* was one of the first attempts to gather a detailed and systematic inventory on an unknown population, for largely unknown reasons, that would ideally have been suited to statistical techniques that did not yet exist.

Despite these limitations, *Domesday* has been described as "probably the most remarkable statistical document in the history of Europe" (McDonald and Snooks 1985; see also Darby 1977). It was so much ahead of its time that neither England nor the rest of the western world would see anything like it for 700 years, during what we could call the second statistical revolution. Here, countries like Norway (1790), the United States (1800), and the United Kingdom (1801) began taking decennial population censuses on every living person.

One of the many things that distinguishes the second revolution from the first was that the state of statistical sciences had increased dramatically. The enlightenment was winding down, yielding a population that was far more literate and numerate, and statistical analysis had burgeoned as a mode of inquiry. Naturally, this stimulated demands for more data, and national inventories had become an essential component of virtually every public policy debate.

The first data revolution was significant for its breadth of measurement, whereas the second stood out for demonstrating how the data could be put to use. In England, once again, people like William Godwin, Thomas Malthus, James Mills, and David Ricardo were at the centre of these debates, sparring over topics as wide-ranging as wealth, health, inequality, free trade, rent, disease, and war, all using statistical evidence as their platform. As only one example of how widespread these debates were, consider that an 1880 bibliography of responses to Malthus's *Essay on the Principle of Population* is over 30 pages long, including only references from the British Isles (Bonar 1970, p. 49).

The second statistical revolution not only gave us an 'avalanche of printed numbers' (Hacking 1982), but it also represented an epistemological shift in how societies see and govern themselves. After the enlightenment, probability rather than indeterminism ruled the day (Hacking 1975), and both sides of any debate were increasingly using statistics to make their case. It was a new way of thinking about ourselves, and it stemmed, at least in part, from the first and second statistical revolutions.

We are currently in the midst of a third revolution, one that departs from, but will likely be as profound as, the first two. In this third wave, which is only a decade or two old, data are collected at a dizzying pace, often without the knowledge, consent, or, it often seems, concern of most individuals. These data are collected by public and private bodies, and have become so large and unwieldy that we can only call describe them with vagaries, like 'big' or 'administrative' data.

As with the first revolution, more data are currently being generated than the society it describes knows what to do with. The near uselessness of much of today's data rivals the uselessness of the number of ploughs in King William's England, because a good deal of 'modern data' has no evident research purpose. If you believe the startling statistic that only 0.5% of all data are currently being analysed (Marr 2015), and that Google alone handles and saves the parameters of 40,000 searches *every second*, it would seem that analysts today are more outmatched by data than ever

before. This would be the case if analytical techniques and computing power and accompanying software were not proceeding apace, but this is not so. Now we have tools like artificial intelligence, deep learning, and big data analytics, to name a few.

That said, it is perhaps the growth in the types of data that are available that is revolutionary. From historical census data to uploaded photos in the cloud, one of the most exciting things about the third revolution is that it will seemingly touch every aspect of our lives. Gone are the days when data were used merely to create theoretical blueprints for understanding society (Condorcet went so far as to predict that death would soon stem only from accidental causes), data today are being used to teach us things that we didn't even know we needed to know.

Early examples of this abound: restaurants knowing where their patrons just were (such as the gym) and adjusting the prices of products (like smoothies) accordingly; using big data to find (or avoid) fellow hipsters; ski resorts using radio-frequency identification tags to cut down on ski lift fraud. We could go on, but you get the idea.

Privacy concerns and potential solutions

As governments become increasingly aware of, and seek to craft policy responses for, privacy implications of the unprecedented extent of data collection on individuals through commercial transactions, social media and cellphone use (e.g., the European General Data Protection Regulation of 2018), developments with government's own administrative data have been in the opposite direction. That is, governments are moving to tap the value for policy of the data they themselves collect through the provision of government services by making those data increasingly available to external researchers. The challenge has been to do so in a way that respects the privacy and confidentiality of the residents on whom the data have been collected. The obligation to do so is almost always enshrined in legislation primarily because unlike commercial data, individuals do not have the option to go without a government service such as healthcare or education nor do they have the option to withhold consent on the collection of those data. However, that legislation may pre-date or may not have kept pace with the demand for—and supply of—administrative data for research.

Approaches to making data more available for research while protecting privacy have included both working within existing legislation and changing legislation specifically. An example of the former is in the U.S. and Canada, where survey, Census and administrative data can be accessed by accredited researchers through secure facilities such as Research Data Centres that exist outside of Government agencies. One recent example of the latter is in the Province of New Brunswick, Canada, where the provincial Government passed enabling legislation specifically to allow administrative data to be shared with a provincial research data centre. Secure data access has also been aided through technological innovations such as virtual private networks (VPNs) that allow remote access to administrative data through a researcher's own computer. A key principle facilitating all of this data access is a refocusing of attention from risk avoidance to risk management, where the benefits of data access are weighed against the consequences of a data breach from the individual's

point of view. Risk of a data breach and the resulting risk of the potential identification of an individual are mitigated through a combination of physical, technological and procedural safeguards plus ongoing monitoring, auditing and reporting. It is those safeguards that have allowed the researchers in this issue to be able to conduct their interesting and policy-relevant work.

About this special issue

Each of the articles in this special issue represent an illustration of what social science researchers can do with their new tools of the trade.

In the paper by McDonald, Liu, and Cruikshank, we see a fascinating example of how health registry data from New Brunswick, a small province in Eastern Canada, can be used to measure immigrant recruitment and retention. As is the case elsewhere, in New Brunswick there is often a significant gap between when a policy is implemented and when researchers are able to evaluate that policy's efficacy. In 2015–2016, New Brunswick admitted more Syrian refugees per capita than anywhere else in Canada. As the province struggles with population decline, one of the hopes is that Syrians will stay in the province over the longer term. Unfortunately, the best data currently available is 3 years old, so it is not possible to identify the secondary migration patterns of recent immigrants. McDonald et al. propose an innovative solution for the data lag in their paper.

In another Canadian paper on internal migration but for Canada as a whole, Haan, Calhoun and Liu examine the likelihood that individuals who left their province of birth return to that province later in life, possibly after exiting the workforce. This is an important question particularly for Provincial Governments already struggling to provide healthcare to their ageing resident populations since a surge in older return migrants will further increase the demands on the Provincial healthcare system. To do this, the authors utilise a unique dataset from Statistics Canada that links 1991 Canadian Census data to annual data on place of residence based on personal income tax returns. Results show wide differences in the propensity to out-migrate, but similarly low propensities to return-migrate, so that people generally are likely to remain in the areas where they spent most of their adult years.

Also studying immigration, Marc-André Luik, Henrik Emilsson, and Pieter Bevelander use administrative data from Sweden to study employment gaps between immigrants and the Swedish-born. They use extremely high-quality register data, and allow them to identify some of the drivers of often-persistent employment gaps between newcomers and the resident population. Admission categories and the transferability of human capital acquired outside of Sweden emerge as important predictors, and the authors argue for programs to help newcomers apply their skills in Sweden.

For Australia, Temple and McDonald study net overseas migration propensities by visa category of immigrants to Australia using a linked dataset developed by the authors that combined international arrivals departures data from the Australian Bureau of Statistics and data on visa grants from the Department of Immigration and Citizenship. The authors are able to examine data on actual

departures from Australia based on passport movements because of the fact that Australia has passport controls for departures as well as arrivals. The authors present estimates of the propensity of individuals granted an entry visa to arrive in Australia as well as the propensity of arrivals subsequently to depart Australia within 16 months of arrival for 22 aggregate visa types as well as 96 specific subclass visa types.

Turning to health, Jatrana, Dayal, Richardson and Blakely turn to mortality differences to showcase how new data allow us to answer old questions. They use census data linked to mortality records to understand the socioeconomic differences in death rates. Given the growing recognition of the importance of the socioeconomic determinants of health outcomes (including death), linkages of this nature are increasingly commonplace. The authors find education to be an important driver of mortality differentials, and encourage future researchers to look at why this link seems to be so important.

The paper by Patler, Sacha and Branic uses administrative data from U.S. Immigration and Customs Enforcement (ICE) on the use of solitary confinement of immigrants in immigration detention facilities, obtained via a Freedom of Information request. This analysis is one of the first published papers on the use of solitary confinement in U.S. immigration detention facilities and represents a major contribution to an important but under-researched subject. The authors find that while female detainees are more likely to be placed in confinement for disciplinary reasons, males are more likely to be placed in confinement for "protective custody" reasons (defined as previous criminal offences, gang status, LGBTW status or other detainee safety reason). As well, mentally ill detainees and those with any medical issue are far more likely to be placed in confinement for reasons related to protective custody than those without medical issues. However, they are far less likely to be placed in confinement for medical reasons. Their analysis also reveals significant differences in the length and reason behind confinement across Field Office Area of Responsibility (AOR), individual facility, and private versus non-private facilities.

Finally, in another paper using linked U.S. data, Foster, Ellis and Fiorio mount a herculean effort to improve migration research in the United States. The authors link multiple datasets from several sources to create what Pia Orrenius has described as the 'holy grail of migration research'. By comparing and validating data from these multiple sources, the authors which will surely send researchers scrambling for ways to gain access to these files.

What each of these articles has in common is an investigation of how new data are changing the way we learn about ourselves. Some argue that things haven't really changed that much, whereas others see the shift as seismic. We hope you find this special issue to be lively contribution to the discussion.

References

Bonar, J. (1970). Introduction. In T. Malthus (Ed.), *An essay on the principles of population*. Harmondsworth: Penguin Books.

Darby, H. C. (1977). *Domesday England*. Cambridge: Cambridge University Press.
Hacking, I. (1975). *The emergence of probability*. Cambridge: Cambridge University Press.
Hacking, I. (1982). Biopower and the avalanche of printed numbers. *Humanities in Society, 5,* 279–295.
Marr, B. (2015). *Big Data: 20 mind-boggling facts everyone must read*. Forbes Magazine. www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#39bd13ff17b1.
McDonald, J., & Snooks, G. (1985). Statistical analysis of Domesday Book (1086). *Journal of the Royal Statistical Society, Series A, 148*(2), 147–160.