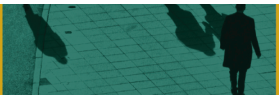
ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS



Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2015)
178, *Part* 4, *pp.* 787–790

Editorial: Big data in social research

Introduction

In the recent American Association for Public Opinion Research report on ‘big data’ (Japec *et al.*, 2015), sources of big data are defined as follows: social media data; personal data (e.g. data from tracking devices); sensor data; transactional data; administrative data. The common feature of these sources of data is that the data are ‘organic’ meaning that they are by-products of processes where the main purpose is not for social research. The sources of data follow the general principles of big data: large volumes of data at high velocity and in varying formats (Laney, 2001, 2012). The authors of the report also include administrative data in the sources of big data. Statistical agencies have long been considering the use of administrative data in their statistical systems and there are many examples of successful applications, particularly in the area of business statistics. Statistical agencies are becoming more involved with processes to improve the quality of administrative data and efforts are made to carry out statistical data editing procedures to satisfy the quality assurance framework of the European statistical systems: relevance, accuracy and reliability, timeliness, coherence and accessibility. The UK Statistics Authority has recently set standards for ensuring the quality of administrative data (UK Statistics Authority, 2015).

In the context of researching social phenomena, we can usefully classify big data into two classes: whether the data records are identifiable, that they can be associated with a single physical unit in space or time, or not. It is this classification which informs how we can use big data for research purposes and what can be achieved. We elaborate on the distinction between identifiable and non-identifiable big data in the following two sections. We outline one method for enabling non-identifiable big data to be used in social research through respondents in a Web panel. We give a few early case-studies of the use of big data in official statistics and close with a brief discussion of privacy and ethical issues.

Identifiable big data

If the elements in a data set can be meaningfully associated with a unit at a given place and time, such as an individual, institution, product or geographical location, then big data can be made fit for purpose for statistical inference. Certainly administrative data would fall under this category. Other examples are satellite imaging for agricultural surveys and censuses, product barcodes from stores to collect data for constructing price indices and traffic loop counters for counting the number of vehicles crossing a specific intersection. Big data may not cover the target population exactly or there may be selective missing patterns in the data that cannot be treated as random, which complicates the statistical modelling and its interpretation. In addition, measurement errors need to be considered when combining sources of data and where the data that are available may only be proxies for the data that are needed. Although these problems are very challenging, statisticians and researchers have developed a wealth of techniques designed to compensate for these issues in their statistical modelling. For example, one way of dealing with these challenges is to ensure that related, high quality, random samples are available. Such random samples can help to compensate for poor coverage, for example, through capture–recapture techniques, or for measurement errors and selectivity when data values are missing.

0964–1998/15/178787

Another notable point is that, when we have identifiable big data, record linkage can be carried out to enhance existing survey and other sources of data. Linkage to administrative data is already an established practice in statistical agencies and is used for enriching statistical data, for example, to carry out small area estimation, or for research purposes to improve the quality of the data collection process. There are new technical challenges in using record linkage with big data, which can be very large and dynamic. These include new methods for blocking, improving linkage algorithms and allowing more transparency so that researchers can compensate for record linkage errors in subsequent statistical modelling.

Non-identifiable big data

Other sources of big data, such as Twitter feeds, other forms of social media and Google searches, require new forms of analytics as well as visualization. This in itself is an important area of research and requires new skills and algorithms. However, if the data cannot be made identifiable at some level then they are of limited use for statistical inference. One example is the classic research of assessing flu epidemics through queries to Google as presented in the 2014 *Significance* lecture by Tim Harford at the Royal Statistical Society international conference (see `http://www.statslife.org.uk/science-technology/1748-tim-harford-and-the-perils-of-big-data`).

Although Google, using an analysis of Web traffic, seemingly predicted well the prevalence of flu at the outset of the relevant epidemic compared with the Centers for Disease Control and Prevention data, it greatly overestimated the trends at later points in time, perhaps because of the intense media coverage driving more frequent searches. This is a good example that illustrates that without knowledge of the causal mechanism, requiring understanding of individual behaviour and hence identifiability of the individuals behind the searches, we need to be very cautious about using such, essentially correlational, data for inference and prediction.

A proposed method for producing identifiable big data

One way to create identifiable big data from social media and search engines is to create a large randomly sampled Web panel. There are numerous examples of Web panels in the USA and at marketing and polling agencies and a large social science Web panel is under current investigation in the UK (see `http://www.cls.ioe.ac.uk/page.aspx?&sitesectionid=1352&sitesectiontitle=UK+Web-based+probability+panel%3a+exploratory+study` for more information about the consultation).

A Web panel is a random sample of individuals who agree to respond to Web surveys that are periodically sent to them, typically via e-mail requests. Although this may be intrusive, one extension is to ask respondents to the Web panel to allow access to their social media activity, such as their Twitter account, which can be monitored and collected for statistical purposes. Under certain *caveats* and agreements with the respondents, it may then be possible to allow statistical inference for this type of big data. As an example, if trends are investigated for election polling, then an agreement could be made with Web panel respondents that their Twitter feeds and social media will be collected if the word ‘election’ is used. An initial exploration of this type of social media data collection is being carried out within the framework of the British Election Study called ‘iBES’ (see `http://www.britishelectionstudy.com/`).

A major challenge for data analysts is to evaluate coverage and selectivity of the respondents for the data to be made useful for statistical inference. We would expect that, as experience with Web panels increases, the usefulness of such panels will increase and this may include the processing of organic data arising from social media for statistical purposes. However, there will presumably always remain a sector of any population who are not represented in such a panel, or who refuse to participate.

Examples of big data applications

Many statistical agencies have devoted resources to investigate how big data can be used in their statistical systems which would reduce costs while compensating for the increasing problem of falling response rates and non-response bias in currently collected statistical data. Below are two examples.

Case-study 1: a task force based on a multidisciplinary consortium from the European Union investigated the use of mobile phone data for tourism statistics (Ahas *et al.*, 2014). The task force provided a comprehensive list of issues that should be dealt with including stocktaking, feasibility of access, methodological issues, coherence of data *versus* collected tourism statistics and opportunities and limitations. The task force concluded that mobile positioning data cannot replace current tourism statistics but can give complementary results. They noted that there are possibilities to explore mixed mode solutions and depending on international co-operation there is a point of referencing for other types of big data.

Case-study 2: research at Statistics Netherlands is outlined in Daas *et al.* (2015). They described two examples: identifiable data in terms of traffic loop detection records to measure traffic intensity at known intersections and non-identifiable data by using social media to assess the sentiment of the population. In the first case, traffic loops can count the number of vehicles per minute that pass at a specific location as well as measure the speed and length of the vehicles. When analysing the data, it was clear that it suffered from selective missing data problems due to some computers failing to submit data. More information about this problem appeared in Puts *et al.* (2015). In the second example, social media messages were extracted that related to sentiment via buzz words. The results seemed to show that social media sentiment across time was highly correlated with the official Dutch Consumer Confidence Survey.

There are many other examples of researchers investigating the use of big data in statistical systems. The Australian Bureau of Statistics is considering the use of satellite data consisting of crop areas at specific time points in their agricultural official statistics production (Tam and Clarke, 2015). In Italy, researchers are investigating the use of big data for small area estimation models (Marchetti *et al.*, 2015). All of these examples indicate that, when big data can be made identifiable, there are potentially large benefits to incorporating the data into statistical systems.

Where do we go from here?

Barriers remain to the widespread use of identifiable data which include the ethical concerns of obtaining consent to the use of identifiable data and preserving the privacy of those whose data are collected. If big data are to be used in statistical inference, then they must be anonymized before being released to researchers to reduce disclosure risk and this will require new skills and information technology solutions.

Clearly, there is a major challenge for record linkage when combining multiple sources of data due to the requirement to respect privacy by reducing the probability of disclosure of sensitive information on individuals or institutions. This involves a balance between reducing disclosure risk by ‘degrading’ data in various ways, while retaining, or being able to recover, sufficient information so that the data are still suitable for efficient statistical analysis.

It is clear that the future of social research is evolving with the emergence of different forms of data, both organic and collected by using structured methods, and the need to incorporate multiple sources of data. Software and algorithms are being developed and the involvement of statisticians in these is essential to ensure that the data that become available retain their integrity and thus usability for statistical analysis.

References

Ahas, R., Armoogum, J., Esko, S., Ilves, M., Karus, E., Madre, J. L., Nurmi, O., Potier, F., Schmucker, D., Sonntag, U. and Tiru, M. (2014) Feasibility study on the use of mobile positioning data for tourism statistics. *Report on Tourism Statistics*. Eurostat, Luxembourg. (Available from `http://ec.europa.eu/eurostat/web/tourism/methodology/projects-and-studies`.)

Daas, P. J. H., Puts, M. J., Buelens, B. and Van den Hurk, P. A. M. (2015) Big Data as a source for official statistics. *J. Off. Statist.*, **31**, 249–262.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C. and Usher, A. (2015) American Association for Public Opinion Research: Task Force Report on Big Data. *Report*. (Available from `https://www.aapor.org/AAPORKentico/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15_b.pdf`.)

Laney, D. (2001) 3-D data management: controlling data volume, velocity and variety. *META Group Research Note*, Feb. 6th. Gartner, Stamford. (Available from `http://gtnr.it/1bKflKH`.)

Laney, D. (2012) *The Importance of 'Big Data': a Definition*. Stamford: Gartner.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L. and Gabrielli, L. (2015) Small area model-based estimators using big data sources. *J. Off. Statist.*, **31**, 263–281.

Puts, M., Daas, P. and De Waal, T. (2015) Finding errors in Big Data. *Significance*, **12**, no. 3, 26–29.

Tam, S. M. and Clarke, F. (2015) Big Data, official statistics and some initiatives of the Australian Bureau of Statistics. *Int. Statist. Rev.*, to be published, doi 10.1111/insr.12105.

UK Statistics Authority (2015) Quality assurance of administrative data: setting the standard. UK Statistics Authority, London. (Available from `http://www.statisticsauthority.gov.uk/assessment/monitoring/administrative-data-and-official-statistics`.)

Natalie Shlomo
University of Manchester

and Harvey Goldstein
University of Bristol
and University College London