Small Steps Towards Big Data

Ric Clarke, Australian Bureau of Statistics
Agenda

Review some basic Big Data concepts
Describe the Big Data opportunity for official statistics
Outline concerns about using Big Data for official statistics
Introduce a framework for statistical inference from Big Data
Provide a snapshot of current Big Data initiatives in the ABS
Fundamental shifts

- Innovation revolution
- The Internet of Everything
- A mobile population
- Ubiquitous connectivity
- The Millennial Generation
- Knowledge circulation
- Service orientation
The ABS view of Big Data

Rich data sets of such **size, complexity and volatility** that it is not possible to leverage their business value with existing data capture, storage, processing, and analysis capabilities.
Key aspects of “bigness”

Size

Complexity

Volatility

Multisource

Multiconnected

Multistructured

Multipurpose

Multidimensional
Big Data, Big Opportunities

Creating sample frames or registers
Providing data for a subgroup of a population
Providing data for some attributes of a population
Enabling data imputation, editing and confrontation
Enabling data linking and fusion
Replacing traditional survey collection entirely
Producing new statistical products
Improving statistical operations
Big Data, Big Challenges

Business benefit
Privacy and public trust
Methodological soundness
Technological feasibility
Data acquisition

ECB Workshop on Using Big Data, 7-8 April 2014
Framework for statistical inference

Target population $U$ on which statistical inferences are to be made
- Example: agricultural land parcels

Big Data population $U_B$ included in the Big Data source, and assume $U_B \subseteq U$
- Example: agricultural land parcels captured in satellite sensor imagery

Measurement $M_U$ of the target population $U$ that is of interest
- Example: crop type associated with an agricultural land parcel

Proxy measurements $Y_B$ (or $Y_U$) available on the population $U_B$ (or $U$)
- Consider $Y_B$ to be a sample (random or otherwise) from $Y_U$
- Example: ground cover reflectance in selected wavelengths

Transformation process $T$ that turns $Y_U$ into $M_U$
- Example: classification model that assigns a crop type to the reflectance measurement of a land parcel

Sampling process $I$ that determines the selection of $Y_B$ from $Y_U$
- Usually unknown and requires detailed contextual knowledge to model

Censoring process $R$ that renders part of $Y_U$ unavailable
- Example: missing imagery data due to bad weather
Framework for statistical inference

For finite population inferences, we are interested in predicting $g(M_U)$ given the observed $Y_B$, where $g$ is some function.

Assume that the probability density function $f(Y_U; \theta)$ is known, where the parameter $\theta$ has known prior distribution $f(\theta)$.

Assume also that the pdf $f(M_U|Y_U; \phi)$, as is the prior distribution $f(\phi)$ of the parameter $\phi$.

Using a Bayesian approach, we want to predict the posterior distribution $f(M_U|Y_B, I, R)$.

In the paper, we show that $f(M_U|Y_B, I, R) \propto f(M_U|Y_B)$ provided that two ignorability conditions are satisfied:

- $f(R|M_U, Y_B, Y_U \setminus Y_B, I, \theta, \phi) = f(R|Y_B)$
- $f(I|M_U, Y_B, Y_U \setminus Y_B, \theta, \phi) = f(I|Y_B)$

i.e. The sampling and censoring processes can be ignored in transforming $Y_B$ to $M_U$. 
Example: sensor data for agricultural statistics

In the case of the remote sensing example, \( M_U \) represents crop type for land parcels in the Australian continent, and \( Y_B \) the remote sensing data covering Australia from Landsat 7.

- As the full data \( Y_U \) is available from Landsat 7, \( Y_U = Y_B \) and the first requirement of ignorability conditions is satisfied.

- When there is missing data, the second requirement needs to be checked. If the missing data is due to random short-term bad weather, it is safe to assume that this is not associated with the measure of interest (crop type) and we treat the data set as a random sample.

- In the case where missing data is due to systemic effects, then an assessment is required to determine whether the observed data set comprises a random sample.
Big Data, Big Technologies

Semantic Web
Machine intelligence
Data visualisation
Distributed computing
ABS use of Big Data

The ABS continually strives to

- Reduce the cost of statistical production and support
- Improve the relevance and timeliness of its products
- Create new statistics that better meet emerging needs

As part of its business transformation program to achieve these aspirations, ABS is taking small steps to exploit particular Big Data opportunities.
Big Data Flagship Project

Build a strong foundation for the mainstream use of Big Data in statistical production

- Methods
- Skills
- Tools and infrastructure

Through a coordinated set of targeted R&D initiatives

- Match Big Data opportunities to specific business problems
- Deliver “fit for purpose” solutions as working prototypes
- Enhance partnerships with academia, industry and other NSIs
- Contribute to a whole-of-government capability
Research areas

Satellite and ground sensor data for agricultural statistics
Mobile positioning data for measuring population mobility
Multiply-linked employer-employee data for productivity analysis
Predictive modelling of survey non-response behaviour
Predictive Modelling of Unemployment for small areas
Data visualisation techniques for exploring large datasets
Sensor Data for Agricultural Statistics

Use of satellite sensor data for producing agricultural statistics

- Landsat 7 imagery from US Geological Survey (multispectral data from 7 frequency bands, 30m grid size)
- Estimate land use and crop type
- Apply machine learning to automated feature recognition
- Promising results for wheat, barley, oats, canola and field peas

Stage 2: extend to the use of ground sensor data

- Sense-T ground sensor data (temperature, moisture, etc)
- Estimate crop yield
- Apply domain-specific agronomic models of crop growth
Questions?