Statisticians or data scientists? The future of official statistics in the era of new technologies and modern data sources

Yoel Finkel

Associate National Statistician, ICBS

Conference of European Statisticians, Geneva

June 2017

***JSSAM* (DEC. 2015)**

Use of big data for production of official statistics POS

Example. 1. Count of number of vehicles crossing road sections. Presently done in a very primitive way.

Why not get the information and **much more** from **cell phone companies?** Available in principle for each point in time.

Exp. 2. Use the **BPP**, based on **5 million commodities** sold **on line** to predict the **CPI→ requires two costly surveys**.

Big Data for POS → Big Problems → Big headache

- High dimensionality and extremely large sample sizes.
- Coverage/selection bias (we are talking of **POS**)
- Data accessibility, new legislation**?** Permission of data subjects?
- Privacy (data protection), disclosure control
- New sampling algorithms
- Data storage
- Heavy computation, new algorithms and analytical tools
- Integration of files from multiple sources at different times
- Risks of data manipulation or sudden unavailability

Do we really get what we need for official statistics??

Two types of big data

Type 1. Data obtained from sensors, cameras, cell phones…, generally structured, accurate, relates to a particular population.

Type 2. Data obtained from social networks, e-commerce…, generally diverse, unstructured and appears irregularly.

- Data from different sources may have different formats, arrive at different times with different degree of reliability, and may be defined differently.

- **No such problems with traditional surveys**.

- **NSOs need to be prepared that data may cease to exist**.

Other important issues concerning big data

Coverage bias- major concern in use of big data for **POS**.

House sales advertised on the internet do not represent properly all house sales. Opinions expressed in **social networks** different from opinions held by **general public**.

- Big isn't always better**!!** Collecting enormous amounts of data **does not guarantee** getting the right answer. A smaller balanced sample may provide better insights than a large skewed one.

No bias when using big data as **predictors** of other variables.

e.g., use **BPP** to predict the **CPI**. Use **job adverts** to predict **employment**. Use **Satellite images** to predict **crops**.

Requires proper statistical analysis to identify and test (routinely) the prediction models.

Other important issues (cont.)

Sampling: random sampling will continue to play a major role in the era of **big data.**

❖ Reduces **storage space**, helps protect **privacy**, produces **manageable data sets** on which algorithms can run to produce **estimates,** and **models** can be fitted.

❖ Sampling from big **dynamic** data **different** from sampling **finite populations**, requiring **new** sampling algorithms; **e.g.,** sampling from social networks **(?)**

❖ If no sampling ⇒ **no sampling errors**. Which measures of error should be computed**? Measure of bias? How? Compare to traditional estimates? Measurement errors?**

Other important issues (cont.)

Big Data for sub-populations: NSOs publish estimates for sub-populations**; age**, **gender**, **ethnicity**, **geography,…**

Big data may not contain this information. Requires **massive linkage if** missing information available in other big files.

Data on sales from supermarkets contains **no information** on buyers. Link to buyers by use of **credit card numbers?**

❖ **Will traditional sample surveys always be needed?**

Other important issues (cont.)

Estimation: at **NSOs** we use **design-based** estimators**, model dependent** estimators**, model-assisted** estimators**,…**

New: ***algorithmic estimators*** **-** the result of computational algorithms applied to the raw **big data** like a complex hierarchical algorithm.

Publication: Big data potentially available for every point in time. **What kind of statistics should be computed and published?** Should official publications from big data be primarily in the form of (online) **graphs and pictures?**

Computer engineering for POS from big data

No longer **Gigabytes** (~10^9 bytes). **Terabytes** (~10^{12}bytes) and **petabytes** (10^{15} bytes) **new standards**.

- Available computing facilities at **NSOs** cannot store and handle such high volumes of data.

Possible solution. Use **cloud** storage, management and processing facilities (Amazon, Microsoft**,…**)

Big problem. **Data protection**. **Many users**, data distributed over a **large number of processors**.

Possible sol. Private cloud (data center). Incorporate **all local computers;** central management of storage space and processing power of separate servers. **Major challenge**.

Big data for POS- summary remarks

New expensive computing facilities, **new** data processing techniques, **new** linkage methods, **new** visualization methods, **new** sampling methods, **new** analytic methods, **new** measures of error, **new** disclosure control procedures, **new** legislation**,…**

Big **potential advantages**: timeliness, much broader coverage (possible **coverage bias**), **no** need for sampling frames, **no** questionnaire, **no** interviewers**,…**

❖ Constant **decline** **in response** rates in traditional surveys and tightened budgets ⇒**use of big data** **inevitable**.

"Good news": Big data will just grow bigger and bigger.

Data accessibility, privacy and confidentiality

- More and bigger pressure by researchers and policy makers to get access to individual confidential data.
- In sharp contrast to promise and responsibility of NSOs to protect privacy.
- If this promise is not held, forget about surveys and censuses.

Two aspects:

A- Cyber security against intruders who may want to:

Disrupt (lock) hardware; Derange or Manipulate data.

- Protection requires **very expensive** hardware and software.

B- Guarantee that data released cannot be used to reveal confidential data. (**Statistical Disclosure Control- SDC**).

"Common" solutions to allow access to confidential data

Research (safe) rooms. Outputs controlled **manually** but more automatic procedures need to be developed in view of increased use by researchers. **Big Data - new major concern**.

Add noise, swap attributes between records with similar characteristics on a set of control variables.

Release "synthetic data" generated from models. **Generate and store new sets of big data? Draw samples?**

Remote access via virtual data enclaves. Enables users to **access the data from their own PC**. Much more resources.

Release only samples of original data. Requires **new sampling methods**, particularly when **sampling from big data**.

Data accessibility, privacy and confidentiality (summary)

- **The more the data is protected- the less is the data utility.**

Requires close cooperation between **computer scientists** developing formal definitions (and algorithms) of disclosure risk, and **statisticians** developing **SDC** methods.

- **Big data** makes everything even more problematic. Much larger volumes of high dimensional complex data, with many more variables and categories than in traditional surveys.

Therefore the future statistician will need:

- Classical statistical theory
- Computer science
- Cyber Security

Other challenges not mentioned in the shortened version with implications for the future official statisticians:

- Mode effects in mixed mode surveys
- Combination of administrative data; record linkage
- Small area estimation

So…are Universities preparing students to work at NSOs?

- **Importance** of **NSOs** **not in doubt**.
- **NSOs** among the largest employers of **statisticians** and **economists**.

Considered three central topics in the work of NSOs**:**

Survey sampling; Seasonal adjustment (SA) and Trend estimation; National accounts.

Browsed the curriculum of undergraduate and graduate courses in statistics and economics of the **top 25 universities** in the world, as ranked by the **Shanghai Academic Ranking of World Universities**.

Are universities teaching the topics in regular courses?

- Only **11** universities offer a course on **survey sampling!!**

  All universities offer a course (**s**) on **Time Series Analysis and Forecasting**, but there are **no specialized courses** on **SA** or **trend estimation**.

  Found few time series courses for which **seasonality** is even mentioned in the curriculum.

- There are no **specialized courses** on **National Accounts**. At best, **NA** mentioned occasionally in **macro-economic** courses, mostly with reference to the **GDP**.

Remarks following these findings

1- What is the role of universities? Are they supposed to train students to work at work places? Should universities focus **solely** on research, and educate new generations of researchers?

❖ The three topics mentioned, and many other topics underlying the work of **NSOs** require **advanced theory**, **no less than in other courses taught regularly**. Teaching courses on these topics is in **no conflict** with the view that universities should concentrate on research.

2- It can be argued that the reason for the lack of courses related to **official statistics** is the lack of **expert researchers** to teach them. **This is the problem**. Students are not exposed to these topics during their studies, and hence they don't even consider them for their academic research.

3- Statistics has changed dramatically in the last decade. Survey sampling and time series analysis have also changed. Courses on such topics need to adapt accordingly.

- **Lohr (Hansen lecture, 2009),** and **Kolenikov *et al*.** (**2015**) discuss in several papers ways of **"Training the modern survey statistician"**.

4- Some good news:

Several universities around the world **do emphasize survey sampling** in their **teaching and research**.

There are several **University Master Programs in Official Statistics** (**JPSM in the U.S., Moffstat in the UK**).

The **National Institute for Statistics and Economics Studies** in France, and the **Brazilian Institute of Geography and Statistics** have schools of statistics with programs for bachelor's and master's degrees in **official statistics**.

The **European Union** has recently established a **European Master Program in Official Statistics (EMOS)**.

Growing recognition of importance of **POS** in **academia**.

Final Recommendation

NSOs should strengthen the ties with the academia by helping establishing specialized courses on POS, and involving more academic researchers in their daily work.

Future use of big data will require new skills from official statisticians. Data scientists, not just experts in statistics or being more specific in survey sampling. Universities and NSOs should be ready to train Data Scientists for official statistics. Some specialized programs in this direction already in planning in several countries.