Daily News

TUESDAY, DECEMBER 3, 2013 DN page vi

'Big Data' for modernizing official statistics

The keynote address delivered by Dr. Amara Satharasinghe, Additional Director General of the Department of Census and Statistics at the second academic sessions of the Institute of Applied Statistics, of Sri Lanka held on November 23, 2013.

It is a great honour and a privilege for me to give at this second Academic Session of the Institute of Applied Statistics, Sri Lanka (formerly known as Applied Statistics Association of Sri Lanka). The Institute can be proud of its, growth and role. I have participated in the previous annual general meetings and I am glad to see the progress of the Institute since its inception. Let me say it is a real pleasure to have this opportunity to speak on the challenges of using 'Big Data' in official statistics as it has been a subject of considerable interest to me for some time now. So let me start by reflecting on the context to these challenges.

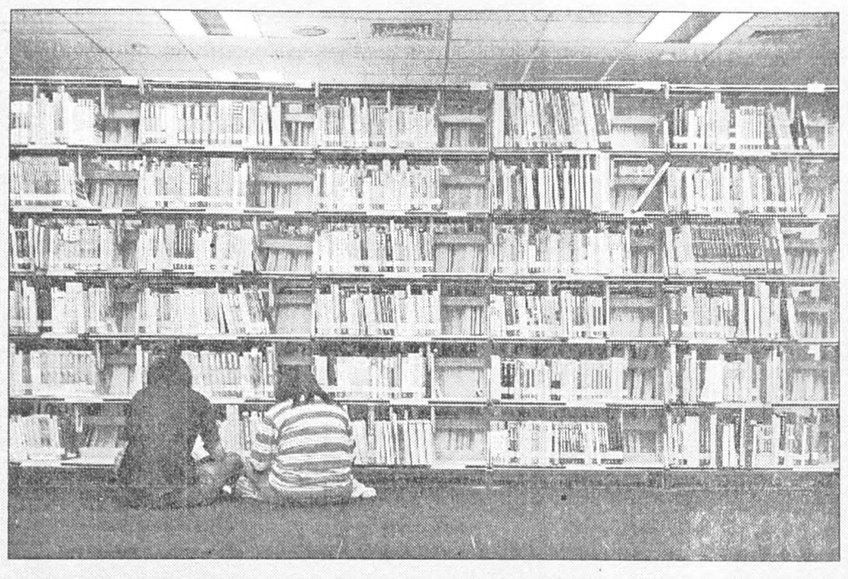
There is an urgency and importance of modernizing official statistics as a strategy for achieving the goal of creating a more adaptive and cost-effective information management environment for National Statistical Offices (NSOs). The key priorities for the modernization of statistical production are the management and implementation of relevant standards, and the development of a "plug and play" approach to producing statistics.

Digital data is streaming in from all sorts of sources; not only from transactional and other business processes, research and administrative systems of both public and private organizations, but from various sensors, instruments, on board computers, simulations and models. Enormous volumes of data are also channeling through communications, transportation, security and logistics networks.

The volume of data and the frequency at which they are produced are so vast that they are usually referred to as 'Big Data'. Big Data is defined as data sources that can be generally described as: "high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making." Big Data are in abundance and reflect part of our evolving way of information heavy living.

Big Data are famously characterized by 3 Vs for their high volume, velocity and variety. Big Data are often also largely unstructured, meaning that they have no predefined data model and/or do not fit well into relational tables. Apart from generating new commercial opportunities in the private sector, Big Data are also potentially very interesting as input for official statistics, either for use on their own, or in combination with more traditional data sources such as sample surveys and administrative registers.

Big Data have the potential to produce more relevant and timely statistics than traditional sources of official statistics. Official statistics has been based almost exclusively on survey data collections and acquisition of administrative data from



government programs,often prerogative of NSOs arising from legislation. The official statistical community needs to better understand the issues, and develop new methods, tools and ideas to make effective use of Big Data sources. This includes closer integration with geographical data and standards to address the issue of how big data can help measure more accurately and in a more timely manner the economic, social and environmental phenomena.

Big Data sources

Traditionally, data processing for analytic purposes followed a fairly static blueprint, with modest amounts of structured data created with stable data models, loaded into an enterprise data warehouse. Non-expert users could then perform basic data visualization and limited analytics via front-end business intelligence tools. With recent developments, data are no longer centralized, highly structured and easily manageable, but are highly distributed, loosely structured and increasingly large in volume. The volume, type and the speed at which new data is created has thus changed and it is also generated by a range of sources, including mobile devices, internet transactions, networked devices and sensors, social networking and media. In general, large data sources can be classified as follows:

a) Administrative: (arising from the administration of a program, be it governmental or not), e.g. electronic medical records, hospital visits, insurance records, bank records, etc.

b) Commercial or transactional: (arising from the transaction between two entities), e.g. credit card transactions, on-line transactions (including from mobile devices), etc.,

Challenges

The use of Big Data in official statistics presents many challenges. Some of the main challenges include:

a) Methodological: Big Data sources register events, not units, and they are selective

b) Methods and models specific for large datasets: Try to 'make big data small' ASAP (noise reduction)

c) Technological:Learn from 'computational statistical' research areas

d) High Performance Computing needs, parallel processing

e) People: Need 'data scientists' (statistical minded people with programming skills that are curious) who are able to think outside the traditional sample survey based paradigm!

National Statistical Offices have started to explore how best to harness this phenomenon of Big Data in their mission to supply quality statistics for improving economic performance, social well-being and environmental sustainability. The attraction lies in the sheer amount of data which could be available in, or near, real time. Potentially, Big Data could be used as intelligence to better solve emergency situations and it also presents an opportunity for the official statistical community to better meet its mission of disseminating timely and quality statistics.

Applications of Big Data

Big Data studies are being conducted or planned in several countries. I would like to elaborate few examples.

i. Traffic and transport statistics: Making full use of this information would result in speedier and more robust statistics on traffic and more detailed information of the traffic of large vehicles which are indicative of changes in economic development. In the Netherlands, about 80 million traffic loop detection records are generated a day. These data are used as a source of information for traffic and transport statistics and -potentially- also for statistics on other economic phenomena.

ii. Real time population density from mobiles: Hourly population dynamics allow urban planners to create better plans for public transport or roadway restrictions and to define spaces of congregation or avoidance.

iii. Mapping Twitter's languages in London Details

iv. Price Statistics: The use and analysis of prices collected on the internet

v. Tourism Statistics: The use of mobile positioning data for tourism statistics.

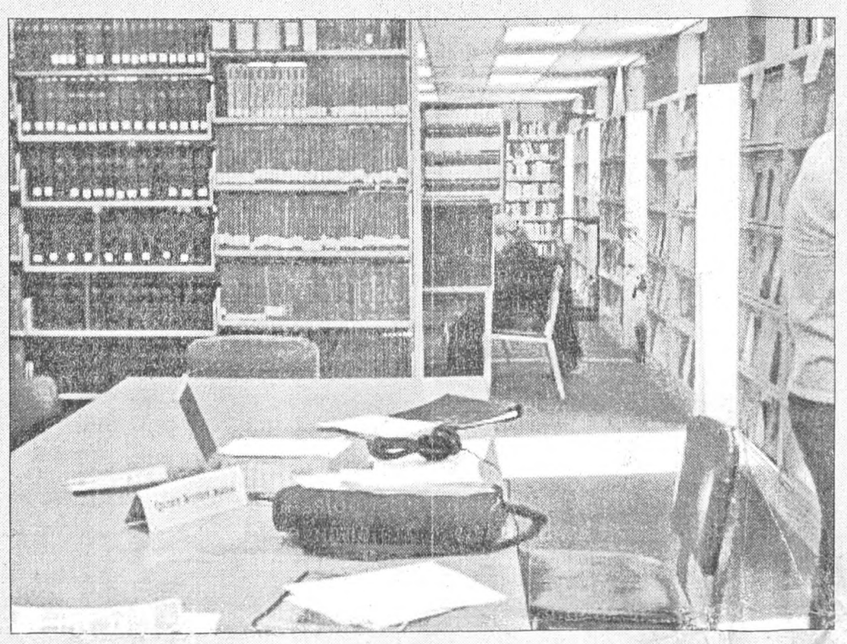
vi. Information and Communications Technology (ICT) usage Statistics: Internet traffic flows for collecting statistics on the Information Society.

vii. Mobile phone data: Travel behavior ('Day time'-population), Tourism (new phones that register to network), Crowd info (for example during events)

The study of phone movements in and between mobile networks enables measuring the flows of people in and between countries, which is why mobile positioning has become useful in different fields of statistics. Mobile telephones are widespread in most countries and they can be used for collecting data for different purposes. For example, mobile data has been used for studying transportation and urban development.

Capacity of NSOs in using Big Data

NSOs are at an early stage of exploring the potential of Big data for purposes of official statistics. But in order for Big Data to truly gain mainstream adoption and achieve its full potential for official statistical purposes, it is critical that the statistical community does not ignore Big Data, but recognizes the use of it as part of their information management model, prepares an inventory of the state of play and formulates the implications for official statistics.



Using enormous amounts of data is not an easy task. Solely because of their size alone, getting insight from Big Data and ensuring quality can be difficult where the data exploration phase would take considerably more time for Big Data compared to other, often more structured, sources of high volume data. As a result,'new' exploration and analysis methods are required. The term new is placed in quotes here because many of the methods exist and are already used but are new in the area for official statistics. Three are found particularly fruitful, namely: Visualization methods, Text mining, and High Performance Computing.

Conclusion

Sooner or later, Big Data will affect the work of national statistical offices in ways—good or bad—that will be primarily be determined by strategic decisions taken in the next few years. Big Data presents significant opportunities and risks for NSOs. The opportunity stems primarily from the so-called "Industrial Revolution of data", underpinned by the spread of digital devices, increasingly hand-held, and characterized by the exponential growth in the volume and variety of high-frequency data—cell-phone data, social media data, transaction data, online news and searches, etc. that capture human actions, experiences, desires, intentions, and expectations.

Big Data are readily available with private companies. As a result, the private sector may take advantage of the Big Data era and produce more and more statistics that attempt to beat official statistics on timeliness and relevance. It is unlikely that NSOs will lose the "official statistics" trademark but they could slowly lose their reputation and relevance unless they get on board. By incorporating relevant Big - Data sources into their official statistics processes NSOs are best positioned to measure their accuracy, ensure the consistency of the whole systems of official statistics and providing interpretation while constantly working on relevance and timeliness. The role and importance of official statistics will thus be protected.

I outlined some of the opportunities and challenges that confront all NSOs as we look to define our place in the rapidly evolving information age that has exploded around us over the past decade. We the statisticians need to position ourselves not just to survive in the digital age but to thrive.

The official statistics community can benefit greatly from the possibilities offered by Big Data, but must invest in research and skills development. Various new areas of expertise are needed to fully exploit the information contained in Big Data. In particular, knowledge is required from the fields of register-based statistics, mining of massive data sets, and the new emerging discipline commonly referred to as 'Data Science'.

We expect to see some official statistics being based on Big Data in the coming years, and are working towards this aim.