Issue Paper 4, Draft ver. 0.1

Big Data and GIS for Disaster-related Statistics

In August 2014, UN Secretary-General Ban Ki-moon asked an independent expert advisory group to develop concrete recommendations to bring about a data revolution in sustainable development. The Report of that group "A World That Counts"[1] describes a multi-pronged strategy to develop and share new technologies and innovations for statistics and to exploit "quick wins" for sustainable development. This includes developing applications of private big data to complement official statistics and to provide enhanced information as a public good with transparency and at the proper scale and in compliance with applicable laws or responsibilities to privacy.

An obvious focal point when considering use of big data for disaster-related goals and targets is the link to geography. Location is crucial to producing disaster-related statistics with increasing usefulness for designing DRR policies. A number of interesting types of big data are becoming accessible for statistics offices that can be used with geo-referencing to improve the availability and level of details for indicators of populations exposed to or directly affected disasters. Examples include: satellite images and other types of remote sensing, records from commercial telecommunications companies (e.g use metadata records from transmission nodes), and consumer data from goods wholesale or resale establishments.

To understand disaster risk and vulnerable groups, there is a need to produce a basic range of baseline statistics on the populations living in areas exposed to natural hazards. At first, this simply means utilizing census data, along with other sources of social-demographic information, as available, focusing on the sample within areas exposed to natural hazards of varying degree, and to the best available knowledge by disaster management agencies (generally, this is detailed utilizing GIS tools in what are commonly called "hazard maps").

GIS creates the possibility for calculation of various new statistics on populations in hazard exposure areas along with integration of any other type of data as long as it is available with geographic referencing at a suitable level of resolution. This function of GIS is currently under-utilized compared to its potential for monitoring sustainable development. Some information could be directly estimated from satellite images. For example, certain kinds of relatively low-cost dwelling structures are more vulnerable to natural hazards simply because of poor durability. Identification of such structures might be observed, or at least estimated, directly from remote sensing. But, more commonly, satellite imagery will be utilized in combination with other types of data at different levels of geographic resolution through a simple grid-based integration approach, for example to model location of people according to the outputs from the population census.

Other sources of geo-referenced data, including data from surveys or administrative records could be utilized to improve the scope for information available on exposed populations, while carefully

[1] http://www.undatarevolution.org/wp-content/uploads/2014/12/A-World-That-Counts2.pdf

protecting individual identities following the usual confidentially-protection measures of national statistics offices (NSOs).

Under the leadership of the Asia and Pacific Expert Group on Disaster-related Statistics, UNESCAP is currently conducting research for leveraging new "big data" sources of geo-referenced information, including data sourced from the private sector. The objective for this research is to develop methodologies that could be applied by national statistics offices, in connection with their existing datasets, to improve the detail, scope, and depth of content of statistics of ex post affected population and *ex ante* exposure to natural hazards.

Some examples for investigation are use of mobile communications and data from wholesale and re-sale on consumer behaviour. This kind of data inherently has a geographic element because mobile phone communication relies on transmission nodes that have specified locations and geographic ranges and wholesale and resale outlets (e.g. grocery or department stores) are immobile and usually cater mostly to proximate populations. Consumer behaviour data, to the extent that it can be accessed for experimental analysis by statistics authorities, might be powerful sources of information to improve or develop more rapid estimates for indicators like poverty or other important social characteristics that are critical factors of vulnerability to disasters and. In addition, theoretically, there may be links to certain changes or alterations in consumer behaviour after or immediately before a disaster and if such relationships could be firmly established though analyses of historical disasters, than theoretically algorithmic models of these relationships could be used to compute rapid estimates of numbers and geographic scope (and perhaps even produce summary statistics on the social demographic characteristics) for the affected population.

Another potentially interesting source of big data source for disaster-related statistics that could be complementary with the idea of telecommunications data are the use information from social networks - e.g. Twitter, Facebook, Baido, Skype, WeChat, etc. These platforms provide not only inexpensive opportunities for instant communication, but also to share data, pictures, and information (e.g. to "check in" their status of “safe” or otherwise affected or unaffected by a disaster). Protocols could be created to give users the opportunity to voluntarily share information and or images, along with GPS location that could be practically utilized by unstructured data analysis programs to support reporting on the extent and magnitude of impacts of a disaster.

Utilizing models or sophisticated data analysis programs for leveraging of big data could help to minimize the extent of impacts in the future and also to improve the quality of statistics as reported from the more traditional sources. In this sense, the relationship between analysis and production of statistics is fluid and sense the baseline exposures to hazards are also dynamic, the models used to extract insightful information from big data will need to be continuously updated and improved. However, in the long-run, use of big data sources could eventually prove to be a more time and cost efficient approach for producing official statistics on disaster risk and impacts.

Working with the large currently available GIS-compatible datasets poses many of the technical challenges that are typical of working with big data and can create a barrier-of-entry for agencies looking to make use of them for official statistics purposes. The files are large and sometimes have limited editing or documentation. There are numerous challenges with integrating across datasets coming from different sources, but through experimentation, standard protocols and workflows can

be developed, for both the production and analysis perspectives, to improve utility of the data for sustainable development analysis