Mobile phone data and geo-spatial information for official statistics and indicators

Ronald Jansen*, Ivo Havinga and Kenneth Iversen
United Nations Statistics Division, New York, United States
jansen1@un.org, havinga@un.org and iversen@un.org

Abstract

The advent of Big Data is having an important impact on the production and analysis of data, and is changing the environment within which the official statistical community operates. Spurred by the increased demands for timely and disaggregated data for the monitoring of the new sustainable development goals of the Post-2015 development agenda, innovative technologies and new data sources are required. The rapid spread of mobile phone usage offers a great opportunity to improve timeliness and fill data gaps. There are today almost 7 billion mobile-cellular subscriptions worldwide. Each of these mobile phones generates data and leaves a digital footprint by communicating, accessing the internet and using various applications. Mobile phone data provides – among others – the location and movement of mobile phones, and is considered one the most promising data sources for measuring and monitoring spatial-temporal activities of the population. In parallel, geospatial information services, including satellite imagery, have added new dimensions to traditional statistical outputs. Together this creates a range of new opportunities to modernize statistical production and improve the quality in different domains of official statistics, including tourism, transportation and population statistics. Privacy concerns and access to mobile phone data are important challenges, as are some technological and methodological difficulties. Despite these challenges, mobile phone data, in particular if combined with satellite imagery and other geospatial information, can offer unprecedented opportunities for improving the timeliness, frequency, coverage and relevance of official statistics, especially in developing countries.

Keywords: Big Data, sustainable development goals, Post-2015 development agenda, satellite imagery

1. Introduction

Big Data encompass a number of different data sources, and have the potential to fill data gaps in a myriad of ways. The emergence of Big Data is radically changing the environment within which the official statistical community operates. At the same time, the prospective monitoring of the Post-2015 development agenda and the Sustainable Development Goals have further put pressure on the official statistical community to adapt and modernize in order to provide better, timelier and more disaggregated data. The Data Revolution called for a monitoring framework that should include all people – leaving no one behind. Data should be disaggregated in ways that allow the relevant differences and similarities between people and groups to be reflected in analysis and policy-making (UN, 2014a).

Big Data can respond to this call by providing data with higher spatial and temporal precision. Compared with traditional sources of data, Big Data can in theory provide real-time estimates, and recent experiments have demonstrated the potential to derive very granular spatial estimates. Different application of Big Data show how it can be used to measure, model, understand and improve the lives of marginalized populations, and provide a new perspective on people for whom quality data have been hard to collect (Blumenstock, 2014).

Remote sensing obtains information about objects from a distance by using satellites, aerial photography and also mobile phones. Data from remote sensing can provide details at high spatio-temporal resolution and are information-rich (UN, 2014b). This paper deals with two separate remote sensing technologies, namely mobile phone data and satellite imagery, and the opportunities these two sources offer for modernizing statistical production.

The spread of mobile phone technology to the hands of billions of individuals might be the single most important change in developing countries since decolonisation and the Green Revolution (UN Global Pulse, 2012). It is estimate that there are almost 7 billion mobile phone subscriptions worldwide, and each consumer contribute to the data deluge with calls, SMS, photos, mobile payment and location traces. The mobile phone network subscription rate globally reached 96 percent in 2014. In developed countries, the number of subscribers has surpassed the total population, with a penetration rate of 121 percent. In developing countries, the subscription rate was 90 percent and continues to quickly increase (ITU, 2014). Call data records (CDRs) can therefore provide information from a wide-range of countries – developed and developing countries alike. In particular, it can deliver detailed information from countries where census and household survey data is non-existent, outdated or unreliable.

Satellite imagery has the potential to complement traditional data sources by collecting data as a relatively unbiased, third party observer. Satellite imagery cannot see currency fluctuations, does not reflect the variable cost of living, and is not dependent on bureaucratic institutions, but can offer insight in a matter of days and detects only real, substantive activity, rather than reflecting accounting values and questionable estimates. Most of the studies employing these technologies have been largely by the academia. However, increasingly, the official statistical community are exploring the opportunities offered to complement traditional household survey and census information. This paper discusses how these two Big Data sources can complement and improve the production of official statistics, in particular in the context of the Sustainable Development Goals. It also discusses the issues and challenges related to using these two Big Data sources for official statistics.

2. A bright example: Luminosity data for economic activity

The idea of using remote sensing technologies for statistical purposes has often been traced back to research, which found that light emissions picked up by satellites can be used as a proxy for standard measure of output. Luminosity is found to track various measures of output, and studies demonstrated that they could supplement national accounting in data-poor countries. There is now a large literature on the use of luminosity as a proxy for population, output, and poverty.

Chen and Nordhaus (2010) compared luminosity and output at the country level for a 16 year period and found that in countries where economic data are extremely weak, luminosity data may be a useful supplement to current indicators as a proxy for output. Similarly, Vernon Henderson et al (2012) used changes in “night lights” as a measure of economic growth, and concluded that satellite night lights data are a useful proxy for economic activity at temporal and geographic scales, in particular in countries where traditional data are of poor quality. In addition, they argued that luminosity data can be useful for all countries to analyze growth at sub- and supranational levels.

3. Make the call: deriving socio-economic indicators from mobile phone data

Building on the experience from the relatively successful luminosity research, and given the ubiquity of cell phones and the information that is embedded in CDRs, a number of studies have evaluated the potential for using this data source for classifying socio-economic status at a granular level. In particular, a number of projects have studied the use of CDRs to estimate or predict income poverty of small areas. By mining mobile phone CDRs, researchers have demonstrated that they are able to predict poverty levels for detailed geographies, which can be used to provide estimates of socio-economic levels that would otherwise not get captured until the next official survey.

Many recent studies have shown the value of mobile phone data to guide development policy and humanitarian action. The Orange D4D challenge, using Cote d'Ivoire as a case study, led to a large number of innovative uses of CDR data, in particular in reference to poverty statistics. If these methodologies, once further refined, reveal promising results, they may provide a way to provide rigorous measures of poverty at very granular geographic levels, at low cost and high frequency, without compromising the privacy of individuals.

In one study, Smith et al. (2013) stressed the need for updated poverty maps, in particular spatially rich and temporally accurate knowledge of socioeconomic indicators. Using mobile phone data for Cote d'Ivoire they discovered a number of features of communication activity that are highly correlated with poverty indicators, and can be used to provide poverty estimates at a fine spatial resolution. In particular, they found that the diversity of a region's connections to other regions can reflect the level of poverty in the region. Similarly, Gutierrez et al. (2013) hypothesized that size and frequency of airtime purchases are correlated with the income of individuals. Despite the lack of up-to-data household income data, they demonstrated that the size of airtime purchases can reflect the user's income or wealth.

Going beyond the monetary measures of poverty Decuyper et al (2015) used mobile phone data to estimate food consumption and food security, by comparing 12 mobile phone variables to 232 food-related indicators. They developed a large correlation matrix which demonstrated that mobile phone expenditure can be used as a proxy for food expenditures in market-dependent households. Based on the mobile phone variables, they were able to create a proxy indicator for the multidimensional poverty index. These findings demonstrate that real-time monitoring of food security could be operationalized in cooperation with mobile carriers. However, most of these studies suffer from the lack of up to date and spatially accurate 'ground-truth' data, which prevent a more rigorous evaluation of the results. In addition, more research is needed to study the potential to disaggregate data by other variables as called for in the Sustainable Development Goals, as well as representativeness and other biases in the data.

4. Moving targets: Tourism and population statistics by mobile phone data

Another strand of research has used mobile phone CDRs to analyze migration patterns. Since census and surveys do not always capture temporal and circular migration, Blumenstock (2012) developed a concept of inferred mobility based on mobile phone data, using a dataset containing mobile phone CDRs for 1.5 million people between 2005 and 2009 in combination with short interviews with 900 subscribers. Similarly, Wesolowski and Eagle (2009) used mobile phone data to better understand the dynamics in the Kibera slum, in particular to study the places of work and migration out of Kibera.

The most advanced project on mobile phone data within the official statistics community analysed the use of mobile positioning data for tourism statistics. Researchers in Estonia have pioneered the use of mobile positioning based methods for studying human movement and tourism since 2001 (Ahas et al, 2011). They relied mainly on passive mobile positioning data in order to compile a number of statistical indicators related to tourism. They found that the inbound roaming data represents the mobility of the foreign tourists within a country very well and have a high correlation with existing data (Tiru, 2014).

Building on the experience from Estonia, Eurostat launched a call for tender for a feasibility study on the use of mobile positioning data for tourism statistics (Eurostat, 2014). The demand for new and more complete data regarding tourism requires that the statistics community rethink how the data is collected. Eurostat seized the momentum from emerging data sources in order to study mobile positioning data for tourism statistics. The study aimed to assess the feasibility of such an approach for producing statistics on domestic, outbound and inbound tourism flows, which clearly corresponds to inbound, outbound roaming and domestic data collected by mobile network operators.

According to the study, the main challenge was the limited access to mobile positioning data due to regulatory limits. The main findings was that mobile positioning data can supplement rather than replace the current official tourism indicators, but that the mobile data introduced several possible improvements in terms of timeliness, access to new information, better resolution, calibration opportunities for existing data, and accuracy in time and space. The study also concluded that other areas of statistics can benefit from joint processes in order to make the processing and use of mobile positioning data more cost-effective.

5. On the map: Mapping populations from space

High resolution, up-to-date data on population distributions are a requirement for accurately measure the impact of population growth, for monitoring changes and for planning interventions. However, the scarcity of mapping resources, lack of reliable data and difficulty in obtaining high resolution up-to-date census statistics remain a major obstacles population mapping in low income countries. WorldPop and Flowminder work to develop new analytical methodologies, using various data sources, to address the critical gaps in mapping resources in order to better understand human mobility. WorldPop have conducted researched that provided high resolution and up-to-date information about human population distributions, and developed a methodology to integrate various data sources to complement information from censuses. WorldPop research also improved the understanding of spatial variation in various socio-economic indicators within countries by developing a novel spatial statistical methodology for the production of gridded surfaces of household survey-based indicators (Tatem et al, forthcoming).

Flowminder focuses mainly on the use of mobile phone data during and in the aftermath of disasters and crisis, in order to fill the gap of basic information on the locations of affected people. Disaster management require accurate information, and mobile phone data have the potential to track population movement in the chaotic aftermath of disasters. Bharti et al (2015) demonstrated how this kind of data can cost-effectively provide detailed maps of population distributions, which is critical for impact assessment and intervention planning in areas such as health, climate change, food security, conflict and natural disasters. This method can also be useful for statistics by providing spatially and temporally detailed dataset, which is an essential denominator for many statistical indicators.

Remote sensing techniques can also be used to improve the sample strategy for household surveys. Munoz and Langeraar (2013), on the basis of Geographic Information System (GIS) techniques, GPS, high resolution satellite imagery, and additional information sources other than censuses, in particular the LandScan population database, design an alternative strategy for sampling design with well-defined selection probabilities. This offers a valuable alternative to the current approach where household surveys use census databases, which are sometimes decades out of date.

6. Bigger than the sum: Exploiting complementarities

Integrating multiple data sources can potentially overcome many of the challenges encountered when using a single source. The combination of different Big Data sources makes the data sources even more useful for estimation and interpolation of socio-economic data. The night-time luminosity studies moved the literature forward by demonstrating the potential of remote sensing data for statistical purposes. However, they encountered some limitations due to the overdependence on night-time lights, which underestimate economic activities that emit less light as they grow. To overcome these limitations, Keola et al (2015) combined night-time lights data with different data sources, in particular the land cover data. This allowed them to estimate growth in areas where night-time lights are not observable from outer space, which is crucial to improve the estimates of agriculturally dominated regions.

Mobile phone data and satellite imagery both have distinct limitations. For example, mobile phone data provide very high spatial and temporal resolution data, but in areas with less cell towers and more irregular use of mobile phones, this data kind of data can be less valuable. Satellite imagery is often limited by environmental conditions such as light contamination and cloud obstruction. In order to overcome the specific limitations of the two data sources, Bharti et al (2015) employed a unique approach of combining these two complementary datasets in order to estimate population movement during the political conflict in Cote d'Ivoire in 2010-2012. This enabled the researchers to measure average long-term population presence as well as dynamic measures of populations across spatial and temporal scales. The researchers claimed that this is the first time mobile phone usage and satellite imagery have been combined to measure populations and movement. Both data sources have in previous studies demonstrated the potential for close to real time monitoring, but encountered different biases and limitations. By combining these two advanced and powerful data sources, the study evaded the limitations of the individual data sources, and extended information beyond the limitations of each data set.

7. Overcoming the challenges: Global cooperation

There are a number of challenges to address in order to provide statistics based on these two sources of remote sensing technologies. In terms of mobile phone data, there are three main types of challenges, namely administrative, technological and methodological aspects. Administrative challenges are related to the access to the data, and the legal and privacy issues associated. Even when anonymized and stripped of all personal information, the data can be used to create highly unique fingerprints of users. However, as demonstrated by several of the above mentioned studies, there are methods that can analyze data aggregated to the level of cell towers rather than individual records, which reduce privacy concerns.

In this spirit, GSMA (2014) developed guidelines on protecting privacy in the use of mobile phone data, in the context of using this data in crisis situations, such as for responding to the Ebola outbreak. This involves anonymization of phone numbers and keeping the data secured and encrypted within the operator's premises. The guidelines document states that all analysis of the data should take place on mobile operator's systems, no analysis that singles out individuals should be undertaken and only the output of the analysis should be made available to the relevant agencies.

Technological challenges refer to the specific data-source as well as scalability of processing performance. The main concern, in particular encountered by many National Statistical Offices is related to the technical ability and resources required to process large amounts of data. Methodological issues are also a major issue with these data sources. In particular, as mentioned by a number of the studies, the representativeness of mobile phone users must be further studied. Even if mobile phone penetration rates are relatively high and are rapidly increasing in most countries in the world, there are still important fractions of the population that are excluded, in particular in rural areas. Therefore a number of caveats must be taken into account in terms of inequalities in access and sample selection bias. Further sensitivity analysis studies are required to analysis the impact of phone use inequality, especially as it can further marginalize already marginalized populations.

To pave the way for more use of mobile phone data the main short term challenge is to improve access to this data source. To achieve this it is necessary to develop mutually beneficial relationships with mobile network operators. One way to achieve this is to establish templates for umbrella agreements on access to data with global companies (UN, 2015).

International cooperation and collaboration is crucial for the development of successful applications of Big Data for official statistics. The Global Working Group on Big Data for Official Statistics, established under the United Nations Statistical Commission, provides strategic vision and promotes practical use of big data sources, capacity-building, training and sharing of experiences. The group are developing pilot studies using both of the Big Data sources discussed in this paper as well as addressing the challenge of access to data and partnership, in particular, investigate the possibility of establishing umbrella agreements on access to data with companies operating globally.

8. Conclusions

Mobile phone data will complement and not replace existing ways of producing official statistics, but the statistical community should have an open mind and recognise the potential of this data source, as demonstrated by the examples discussed in this paper. In particular, mixed-mode solutions, that are large samples based on Big Data, combined with smaller surveys, should be further explored. One first step is to join forces across statistical domains that can make use of the same Big Data source. As demonstrated above, mobile phone data, in combination with satellite imagery can be used to estimate a number of statistical indicators ranging from poverty, food security, migration, mobility, tourism and population.

References

Ahas R, Tiru M, Saluveer E, Demunter C (2011) ‘Mobile telephones and mobile positioning data as source for statistics: Estonian experiences’.

Bharti, N, Lu, X, Bengtsson, L, Wetter, E, and Tatem AJ (2015) ‘Remotely measuring populations during a crisis by overlaying two data sources’, International Health, 7 (2): 90-98.

Blumenstock, J. (2012) ‘Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda’. Information Technology for Development. 18 (2). Pg 107-125.

Blumenstock J (2014) ‘Calling for better measurement: Estimating an individual’s wealth and well-being from mobile phone transaction records’. Proceedings of Knowledge Discovery in Data.

Chen and Nordhaus (2010) ‘The value of luminosity data as a proxy for economic statistics’, NBER working paper series, Working paper 16317, August 2010.

Decuyper, A, Rutherford A, Wadhwa A, Bauer JM, Krings G, Gutierrez T, Blondel VD, Luengo-Oroz M (2015) ‘Estimating Food Consumption and Poverty indices with Mobile Phone Data’. arXiv:1412.2595v1 [cs.CY] 22 Nov 2014

Eurostat (2014), ‘Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics’. Luxembourg: Publications Office of the European Union, 2014.

GSMA (2014) ‘GSMA guidelines on the protection of privacy in the use of mobile phone data for responding to the Ebola outbreak’.

Gutierrez T, Krings G, and Blondel VD (2013) ‘Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets’. arXiv preprint arXiv:1309.4496 (2013).

ITU (2014) ‘Measuring the Information Society Report 2014’ (PDF). Geneva, Switzerland: International Telecommunication Union (ITU). 2014.

Keola S, Andersson M and Hall O (2015) ‘Monitoring Economic Development from Space: Using Nighttime Light and Land Cover Data to Measure Economic Growth’, World Development Vol. 66, pp. 322–334, 2015.

Munoz J and Langeraar E (2013) ‘A census independent sampling strategy for a household survey in Myanmar’. Available at: bit.ly/TU94rr.

Smith C, Mashhadi A and Capra L (2013) ‘Ubiquitous Sensing for Mapping Poverty in Developing Countries’. Paper submitted to the Orange D4D Challenge.

Tatem A, Gething P, Pezzulo C, Weiss D, Bhatt S (forthcoming) ‘Development of Pilot High-Resolution Gridded Poverty Surfaces: Methods working paper – poverty’.

Tiru M (2014) ‘Overview of the sources and challenges of mobile positioning data for statistics’, Background paper prepared for the International Conference on Big Data for Official Statistics, Beijing, October 2014.

UN (2014a), United Nations Secretary-General’s Independent Expert Advisory Group on a Data Revolution for Sustainable Development (2014) ‘A World that Counts Mobilising the Data Revolution for sustainable Development’, November 2014 http://www.undatarevolution.org/report/

UN (2014b) ‘Prototype Global Sustainable Development Report’. Online unedited edition. New York: United Nations Department of Economic and Social Affairs, Division for Sustainable Development, 1 July 2014. https://sustainabledevelopment.un.org/globalsdreport/

UN (2015) ‘Report of the Global Working Group on Big Data for Official Statistics’, E/CN.3/2015/4.

UN Global Pulse (2012) ‘Big Data for Development: Challenges and Opportunities’. New York: UN Global Pulse. May 2012.

Vernon Henderson J, Storeygard A & Weil D (2012). ‘Measuring Economic Growth from Outer Space’, American Economic Review, American Economic Association, vol. 102(2), pages 994-1028.

Wesolowski A and Eagle N (2009) ‘Inferring Human Dynamics in Slums Using Mobile Phone Data’. Technical Report, Santa Fe Institute.