Round table on Web and Social Media for Demographic Research

Foz do Iguaçu, Brazil, 19 October 2016

Chair: Maria Martha Santillan (CIECS, CONICET & UNC – Argentina)
Discussant: Julio Ortega (Universidad San Francisco de Quito (USFQ) – Ecuador)

1. Demography-driven Discovery in the 21st Century: Challenges, Opportunities and Ambitions for a Wonderful Discipline. Emilio Zagheni (University of Washington – United States), présentée par Julio Ortega
2. Production of experimental statistical information from data available on the web, using Big Data technologies. Gerardo Leyva (Instituto Nacional de Estadística, Geografía e Informática (INEGI) – Mexico)
3. NSOs and NGOs: Talkin' bout a (geo) data revolution. Javier Carranza (Geo-Censos – Columbia)
4. Online Data Noisy: Challenges and Implications. Fabricio Benvenuto (Federal University of Minas Gerais – Brazil)
5. The Data Revolution: IUSSP Activities and the Role of Demographers. Tom LeGrand (IUSSP Vice President and Université de Montréal – Canada)

Each panelist was given 10 minutes to present their topic and, at the end of the presentations and the discussant remarks, there was a period of questions and answers from the floor. The main points of the presentations are described below:

1. Julio Ortega, who presented on behalf of Emilio Zagheni, spoke of an integrative approach between traditional and new forms of data collection and described how data analysis is evolving due to rapid technological change noting the research areas in which these new data can be applied.
2. Gerardo Leyva spoke about the complexity of using social media data. His presentation started with a definition of Big Data and then examined the capture of geo-referenced tweets revealing the degree of infrastructure required for their analysis drawing on the work of the Instituto Nacional de Estadística, Geografía e Informática (INEGI – Mexico) in this area. This presentation exemplifies the approach proposed by Emilio Zagheni providing concrete examples, such as using new data on tourist flows to examine human mobility in real-time. The presentation included the use of Machine Learning algorithms to validate data. It also addressed subjective sentiment analysis of messages sent through tweets about specific news items, such as positive or negative perceptions voters have of Clinton and Trump. Two questions arise concerning the use of these data in Mexico: What percentage of Mexico's population uses Twitter? And, of those who use Twitter, what percentage can be geo-referenced?
3. Fabricio Benvenuto: One of the main concerns in using new types of information is assessing the validity of the data and looking for possible bias. This presentation provided an idea of how to "clean" these data. It also shows the value of Twitter to quickly assess and deal with natural disasters (e.g., the Japanese earthquake), diseases (e.g., the dengue epidemic), responses to elections, the analysis of human migrations, etc. The scope and limits of online data, as well as its credibility, were examined, indicating what is possible and not possible to seriously study with data from these new sources. An extreme case of assessing the validity of these data concerns the need to detect if tweets are written by “real” humans as opposed to

computer “bots”. The presentation suggested that data mining methods will be useful in the analysis of new data, as they have been with traditional data.

4. Javier Carranza demonstrated the use of "open" geodata to create new options for analysis and collaboration. Valuable data are those that can be used freely by citizens and which are validated by the government, such as for example OpenStreetMap. Geo-Censos promotes the generation of geodata from civil society to encourage its use by civil society, so as not to depend entirely on official data.
5. Tom LeGrand provided an overview of the Data Revolution: the rapid growth of the new data and their different formats (images, texts, etc.), and the need for new approaches and methods to effectively use them. In addition, implementing the United Nations 2030 development agenda requires more rapidly measuring the SDGs at disaggregated levels, drawing upon both traditional and new types of data. The IUSSP is working to promote the place of demographers in these developments, as they bring with them valuable methodological tools and the ability to assess the usefulness and limitations of new (and traditional) types of data. There is a need to devise new ways to validate and calibrate new types of data so as to avoid biases and provide confidence intervals. Demographers have much to contribute and also much to gain by being more involved in this work. Indeed, these developments may be of use to draw into demography students with interests in mathematics or science computer, to the benefit of all parties.

   The following is a bullet point list of some of the ideas that were discussed about Big Data Revolution, some of which remain as questions:

   - Recognize the use and limitations of new types of data
   - Interoperability: the value of linking data from different sources
   - Confidentiality
   - Representativeness
   - Access to the data
   - The gaps of knowledge concerning the use of Big Data
   - How to improve the capacity of National Statistical Organizations to access and use this information
   - How new students enrolled in demography and related field study programs can be encouraged by the use of these new approaches to data: cellphones, satellites, twitter...
   - The measurement of the SDGs requires the use of new approaches in the type of data.

   These new types of data are:

   - Free (no cost)
   - Generated in real time
   - Subject to biases
   - Validation can make use of Machine Learning
   - Require ongoing calibration

   Challenges:

   The spread of the use of these new types of data, the training required for its use, the confidentiality of the information, the validation of the data, the comparability with the information that is obtained from traditional data.

Conclusion:

As these new types of data become available, new technological tools are being created to improve their reliability (Machine Learning, for example), to remove or minimize biases and to increase the usefulness of this data for multiple purposes and objectives; not only by demographers but also for human knowledge in general.

The idea is to combine the traditional sources of data with the "new" sources in order to obtain information that could be used in every field of knowledge. Many of the concerns over this very rapid expansion of data and methods are comprehensible, in part due to fear of the unknown.