Using Big Data to Supplement Official Statistics

Big Data has become a very popular term in the past decade, but mostly in the realms of Computer Science (known as Data Mining) and Marketing (known as Analytics). This exploratory study looks at how Social Scientists and Demographers are starting to incorporate Big Data into their analyses and the many possibilities that are yet to be exploited.

Just as astronomy and biology underwent revolutions with the invention of the telescope and the microscope, the social sciences are undergoing a similar transformation. Social researchers who continue to go out into the field to measure their concepts by hand are bound to be left behind by those who are able to obtain and organize vast amounts of unstructured data and explore whatever interesting patterns they can find. In a way, this is a shift from using deductive to inductive methods in social research.

Big Data has enormous potential in the developing world, where official statistics are normally not as available or reliable but where mobile phones have an increasingly large penetration. By analyzing this data, for example, the UN Global Pulse wants to try to predict economic crises, hunger and other problems that were previously left unrecorded. I give an overview of many other possible benefits that Big Data could bring to social scientists.

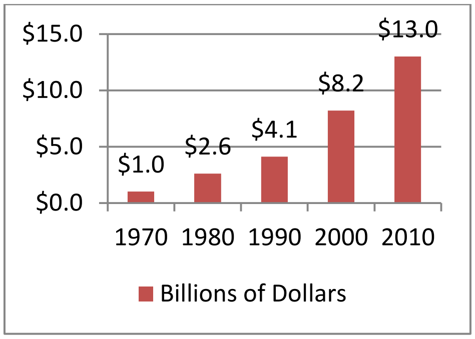
At the same time, I also analyze the potential risks of big data and how these could be mitigated. Privacy is one of the main concerns before being able to use big data. However, if companies that gather big data strive to transform themselves into “data donors” and show clear benefits of how the data is being put to good use, people might be more willing to share their information.

The study also examines the reasons for the current deficit of these tools in demography and how this could possibly be addressed. The tools currently used to analyze large datasets are taught in Computer Science departments. However, most of these computer scientists do not know much about the social world and thus do not know what questions to ask in the first place. However, quantitative social science programs are not that far off from undergraduate programs in computational science, and offering these courses as electives could go a long way to give these much needed skills to the social scientists of the future.

Additionally, the paper looks at how the Census Bureau is spending too much, compared with other countries, collecting data that could be obtained for a much smaller price. European countries using a population register, especially Denmark, Netherlands and Finland, have a cost of less than one dollar per person they enumerate. The United States, on the other hand, spends more than forty dollars per person.

Since the United States has above average telecommunications and many digital transactions, there is plenty of data that could be used, in conjunction with official statistics, to obtain more accurate data. Specifically, credit card information, which has addresses and is verified to be true, could be used to obtain personal information. I argue that a population register, supplemented by big data, seems the most appropriate way to reduce the cost of enumeration while obtaining more timely and useful results.

Figure 1. Evolving Cost of the US Census



Finally, there are many legal barriers to trading in data. Private companies own the data that users give them freely but the government seems to now want a part of the data. Academics could possibly stand between businesses and government, ensuring transparency, confidentiality and promising to obtain information that will be used for the social good.