Big (Geo)Data in Social Sciences: Challenges and Opportunities

Javier Gutiérrez-Puebla
Universidad Complutense de Madrid
javiergutierrez@ghis.ucm.es
Juan Carlos García-Palomares
Universidad Complutense de Madrid
jcgarcia@ghis.ucm.es
María Henar Salas-Olmedo
Universidad Complutense de Madrid
mariahenar.salas@pdi.ucm.es

Keywords: Big Data, geolocalized data, social networks, mobile telephones, smart cards

The production of geographic information is proceeding at a pace that was previously unimaginable. Institutions that do not make their geographic information available to the public, either through download areas on their websites, Open Data platforms or the creation of Spatial Data Infrastructures (SDIs), are few and far between. Companies are also eager to join in the production of massive data, with the aim of improving their productive processes and competitiveness or opening up new niche markets by offering new information and services. However, it is possible that the greatest acceleration of the production of massive geolocalized data has been brought about by voluntary actions, supported by the development of Web 2.0. One of the best known examples of this is Open Street Maps, a collaborative project consisting of the creation of a digital street plan with worldwide coverage that is drawn up and constantly updated by volunteers. The street plan can be downloaded free of charge on the internet.

Big Data is an emerging concept that has become hugely popular over the last two or three years. It refers to the production of enormous quantities of data through multiple networks of sensors and devices. It is not only the massiveness of the data that should be emphasized but also the fact that they are of a different nature to conventional data and, as such, are complementary to data provided by official statistics. In the current technological age, human activities, either voluntary or involuntary, leave a digital footprint that is frequently geolocalized. For example, we generate huge quantities of geolocalized data when we travel (registered by the GPS on our smartphones), when we make a mobile phone call, when we pay by credit card or present a loyalty card, when we interact with social networks or when we are

Revista de Estudios Andaluces, vol. 33, núm. 1 (2016) pp. 1-23. e-ISSN: 2340-2776
http://dx.doi.org/10.12795/rea.2016.i33.01

caught on video cameras in a shopping mall. As Batty (2013) shows, production of the greater part of the information we now call Big Data is both automatic and routine and uses different types of sensors. In principle, the object of almost all this data gathering is to carry out control and management processes in companies (for example, the management of credit card charges), but data have also been used for purposes that are different from those for which they were stored, such as consumer behaviour analysis for designing marketing strategies, the prediction of market tendencies, fraud control, or the generation of new updated and more detailed statistics (Heershap et al., 2014). A well-known case in the field of marketing strategies is that of Facebook, which uses algorithms based on user behaviour on the network (for example: likes, content visited) to create a profile of each user in order to send them personalized publicity.

In a Web 2.0 world, internet users are no longer mere passive receivers of information; they become producers of enormous quantities of data, particularly through social networks. The geolocalization of tweets or the analysis of relationships through Facebook are two of the best-known examples that offer great possibilities for analysing social networks and their spatial footprint.

Several recent studies published in the Dialogues in Human Geography journal focus on the challenges and opportunities of the use of Big (Geo)Data in Human Geography in general, and on city and urban planning studies in particular (see Batty, 2013; Kitchin, 2013; Graham and Shelton, 2013). As the title of the work by Graham and Shelton (2013) suggests – "Geography and the Future of Big Data, Big Data and the Future of Geography" – the future of geography and massive data go hand in hand. The sources of data available enable researchers to end their dependency on official statistics in such diverse fields as demography, economic activity, mobility or any other urban aspect (Shelton et al., 2015). It is therefore not surprising that these new data sources have triggered a growing interest in geographers, who are increasingly attracted by the innovative character of the data and the relevance these are beginning to acquire in studies of the modern day city.

Among the main advantages of Big Data are high spatial and temporal resolution, allowing for the monitoring of spatio-temporal processes, and the possibility of offering information that is complementary to that from official sources. Because a large part of Big Data contains geolocalized data, this can be treated and analysed with Geographic Information Systems and spatial statistical techniques, as well as with classic statistical tools. Animated visualization takes on a particular importance in this type of geolocalized data, precisely because of its high spatial and temporal resolution. In short, massive data can be processed and converted into information from which knowledge can be generated.

Although there is no clear consensus in this regard, the main characteristics of Big Data are cited below (Kitchin, 2013):

- Its enormous volume (terabytes or petabytes of data).
- The great speed at which it is generated (in real or almost real time).
- Its diversity and variety, including data that is structured (in database format) and unstructured (for example, lines of text).
- Its comprehensive coverage of entire populations or systems.
- Its fine-grained resolution, with identifiers allowing individuals or objects to be followed.
- Its relational nature, as it has common fields that allow different databases to be combined.
- Its flexibility, with respect to both its extension (new fields can easily be added) and its scalability (the volume of data can be increased).

Big (Geo)Data covers very different spheres: mobile phone activity, registration of credit card transactions, data gathered in real time with GPS, social networks, registration of water and electricity consumption, images recorded with video cameras, public transport cards or public bicycle systems, etc. It also has a great variety of research applications in fields such as marketing, mobility, tourism and social differentiation. This study reviews the different works of research that have used Big Data, ordered according to the sources used. It is not a complete review but it provides an initial approach for the researcher entering the world of Big (Geo)Data for the first time.

The use of Big (Geo)Data in social science research is only in its early stages but there is no doubt that it is opening up new and promising possibilities. Not only does it become possible to answer some traditional questions from different perspectives (as a result of the greater spatial and temporal resolution of the data, for example), but it allows research questions to be formulated that could not be answered using traditional sources.

The use of Big Data has great advantages for the researcher. It provides information that complements that of traditional data sources, allowing responses to research questions from a different perspective. Many of these sources have global coverage, making it possible to carry out comparative studies between cities or countries. Also, massive geolocalized data have a very high spatial and temporal resolution: spatial, because each item of data is localized by its geographic coordinates and not spatially aggregated; temporal, because the moment of generation of each item of data is stored (year, month, day, hour, minute and second), so that the data available is always up-to-date, allowing evolving studies to be carried out and processes to be

http://dx.doi.org/10.12795/rea.2016.i33.01

monitored. Some of these data sources are free and can be downloaded directly from the internet using the corresponding API.

However, not everything is an advantage. Some of these massive data sources are difficult to access, either because the companies that generate them are unwilling to share them with researchers, or because they charge high fees. A second problem arises from the difficulty of processing massive data as their volume exceeds the capacity of conventional database managers. A final drawback is that such data are generally biased, which means they have to be complemented by other sources in an attempt to compensate for this. Consider, for example, the use of social networks, which are not used by the entire population and within each user group the intensity of use can be quite different. The corollary of this is that analyses carried out with Big (Geo)Data are fundamentally exploratory in nature and a statistical determination of the margins of error and significance levels is generally not possible.

http://dx.doi.org/10.12795/rea.2016.i33.01