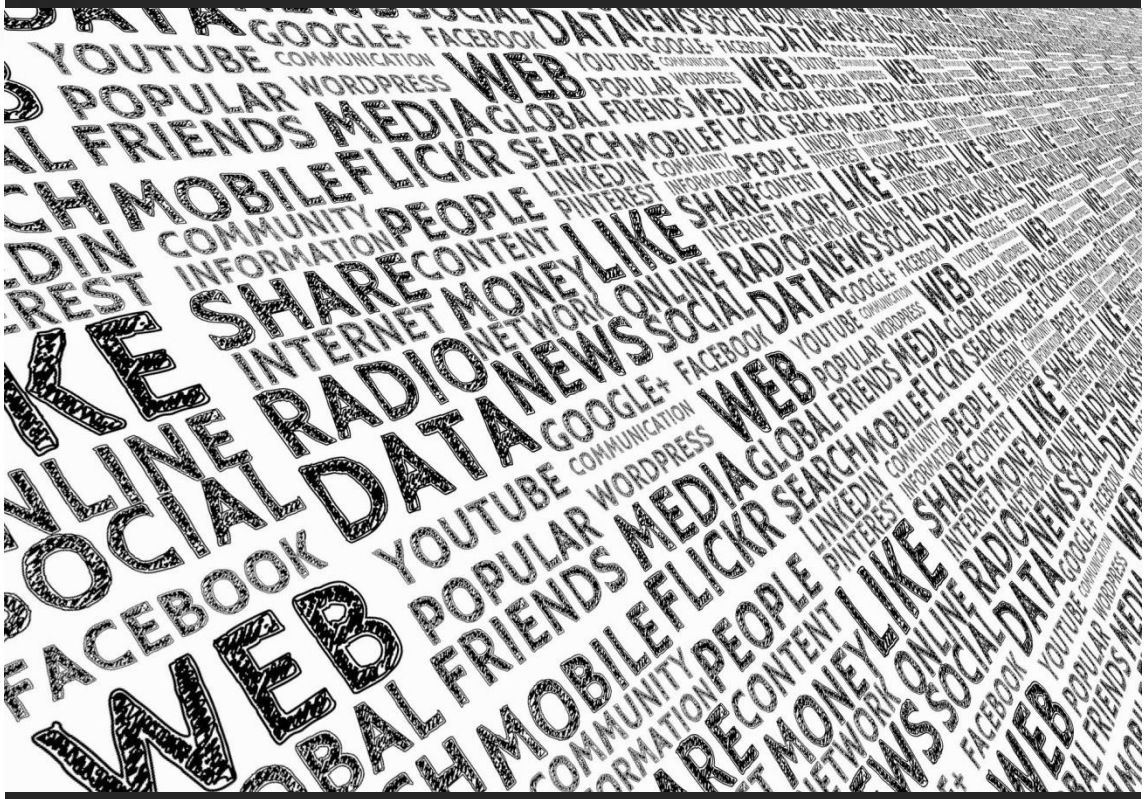
Research Workshop on
Digital Demography in the Era of Big Data
Seville, 6-7 June 2019



Pabellón Nueva Zelanda.
C/Leonardo Da Vinci, 21. Isla de la Cartuja. 41092
Sevilla, Spain

Instituto de Estadística y Cartografía de Andalucía
CONSEJERÍA DE ECONOMÍA, CONOCIMIENTO, EMPRESAS Y UNIVERSIDAD

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

CCHS

iegd

LONGPOP

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676060

MAX-PLANCK-INSTITUT FÜR DEMOGRAFISCHE FORSCHUNG
MAX PLANCK INSTITUTE FOR DEMOGRAPHIC RESEARCH

Fundación BBVA

TELNET
Redes Inteligentes

Summary Report

The Research Workshop was organised by the IUSSP Digital Demography Panel, LONGPOP H2020 Marie Sklodowska-Curie ITN project, Max Planck Institute for Demographic Research (MPIDR), DisCont (ERC Advanced Grant, Bocconi University), the Institute of Economy, Geography and Demography (Spanish National Research Council - CSIC) and the Institute of Statistics and Cartography of Andalusia (IECA). The workshop took place at the IECA in Seville, and brought together 30 scholars to discuss the implications of digital technologies for demographic behaviour as well as the applications of new data from digital sources to understand population processes. Sixteen papers were presented, including 2 invited presentations.

The workshop was preceded by a pre-meeting that included a Demography Today Lecture by John Palmer from the Pompeu Fabra University on *Digital Demography, Human-Mosquito Interactions, and the Socio-Ecological Context of Vector-Borne Disease*, and a training session by Emilio Zagheni from the MPIDR on *Accessing and making sense of digital trace data for demographic research*, both of which were sponsored by the BBVA Foundation.

Day 1: 6 June

The Research Workshop was officially opened by Elena Manzanera, Director of IECA, Juan del Ojo, Subdirector of the Area of Coordination, Communication and Methods at IECA, and Giampaolo Lanzieri, Senior Expert at the Statistical Office of the European Union (Eurostat). Participants were then welcomed by Emilio Zagheni (MPIDR) and Francesco Billari (Bocconi University), the two Chairs of the IUSSP Panel on Digital Demography, and by Diego Ramiro Fariñas (CSIC), member of the Panel's steering committee and local organizer of the workshop.

The first session, on Digital Demography, included two presentations. Samin Aref (MPIDR) introduced his work using the Web of Science (WoS) to track the international mobility of researchers through their affiliations in the publications. In this project, based on over 200 million authorships (1956-2016), the authors have exploited the linkage between an author's name and a publication. Despite the limitations encountered, such as time variation or that not all movements result in a publication and not all the publications are indexed in the WoS, they have discovered some patterns and common destinations for researchers. Sofía Gil-Clavel (MPIDR) made a presentation on demographic differentials in Facebook usage around the World, using disaggregated data from 136 countries by age and gender, retrieved from the Facebook Marketing Application Programming Interface (API). Some results revealed that the greatest median gender disparity in favour of the male population is in Asia, while females tend to be more engaged with Facebook in North America, and that away from their hometown, women are more likely than men to engage with the Facebook platform.

The second session brought different methods to study mobility and migration. Dilek Yildiz (Wittgenstein Centre for Demography and Global Human Capital) proposed a Bayesian probabilistic hierarchical model for combining traditional media (such as Eurostat, EU Labour Force Survey, population and housing census) and social media (Facebook) migration data between 2011 and 2018. The aim of their study was to provide an estimate of true working age bilateral EU mobility stocks with uncertainty. Different data/countries provide higher or lower credible intervals due to the data sources available. Asli Ebru (Bocconi University) presented analyses of the potential predictive power of Google search data (i.e. Google search queries for province names) to observe the movement of Syrian refugees under temporary protection in Turkey, across provinces. Also, they studied the time-lag between a potential change in Google

search data and the actual change that appears on official records. Some of the preliminary results showed that search queries made in Syria for province names in Turkish were strongly significant (but not those in Arabic), as well as a positive and significant effect between search frequency in Turkey and number of Syrian refugees under the temporary protection status.

The third session included two case studies which complemented the previous session on migration: Latin-Americans in Spain and the Indian and sub-Saharan diasporas. By exploiting the digital footprint by potential migrants in Google (Google Searches indexes in Argentina, Colombia and Venezuela between 2007 and 2016), Juan Galeano (Centre d'Estudis Demogràfics) investigated if it was possible to predict their entry in Spain. Findings demonstrated that the analysis of correlation between entries and search indexes can be very useful for predicting trends. But while in the case of Venezuelans, the search index of Google Trends contributed slightly to better predict data, in the case of Argentina and Colombia, the addition of the index entails an underestimation of predicted flows in relation to the base model. Nachatter Singh (Centre d'Estudis Demogràfics) also used Facebook to understand the mobility of highly educated immigrants by gender from Indian and sub-Saharan African diasporas, in comparison with the United Nations Global Migration Database. Results revealed some advantages, such as access, free of cost, to a considerable of information on the socio-demographic characteristics of active users, but which could be user-induced misinformation and limited in terms of age. The main issues raised were related to Facebook data in the countries where Facebook is forbidden or is being used less now. In case of both, one of the suggestions was to look at the returning migrants too.

The last session of the day focused on poverty and energy. Jordi Ripoll (Devstat, Spain) presented his work on exploring the use of an e-commerce dataset to measure poverty levels in Brazil, linked with the official statistics data from the country. Findings were positive when a smoothed ratio was used because it considers the spatial correlation observed in the purchases, contrary to the raw ratio of purchases, which had an excess of zeros due to the absence of observations in 27% of municipalities. This work was followed by the presentation by Vasileios Giagloglou (TELNET, Spain), in which he introduced the work carried out by Energy Minus+ in Machine learning with electrical data to predict the behaviour of a system related to external variables in order to: detect anomalies, predict savings and confirm and validate savings. He also presented an overview of his work for the H2020 LONGPOP project using Elasticsearch to harmonise databases for research.

Day 2: 7 June

Guangqing Chi (Pennsylvania State University) presented an overview on how to retrieve data on migration, for example by comparingmigration estimates from tax return files by U.S. Internal Revenue Service – IRS) with Twitter data, as well as the challenges of using Twitter for this purpose. The data used in this study had been collected since 2013 and he displayed tables and maps comparing IRS data with Twitter data. Findings showed that there were some places in which Twitter is able to predict IRS, but others where it was not possible, and migration flows were under- or over-represented depending on the area.

By moving to the use of mobile phones for the study of demography, Valentina Rotondi (Bocconi University) and colleagues provided large-scale evidence that mobile phones can be a vehicle for sustainable development. Data was retrieved from many different sources. For example, global macro-level evidence using the UN gender inequality index correlated with mobile phone subscription/population revealed that with more penetration of mobile phones, there

was less inequality. Despite the literature criticising the use of mobile phones (individualism, isolation, etc.), there were positive effects in terms of reducing child and maternal mortality, narrowing gender inequalities and enhancing contraceptive use. However, there are still gender gaps in access to mobile phones and in use, so if these two gaps are not tackled, the potential of mobile for development is still reduced, according to Rotondi.

This was followed by two presentations on mobile phones and population estimation. Romain Avouac (ENSAE ParisTech) gave insights on the use of a Bayesian approach to improve the estimation of population using mobile phone data. The main contribution of their study was the improvement of spatial mapping through combination of data sources and the use of a modular approach. Thanks to this, it was possible to detect homes and statistically adjust (correct for heterogeneous market shares and penetration rate) and to compare with population estimates from tax data. This system had an important effect in rural areas, but less in densely populated areas such as Paris. In the second paper, Fabio Ricciato and Giampaolo Lanzieri (Eurostat) proposed a methodological framework for estimating present population density from mobile network operator data. Their research identified some challenges such as the variability of the technology used (2G/3G/4G), the complexity of the extraction of the most/best information from raw data or the estimation of the mobile phones per person (double counting, under-coverage or over-coverage). They also proposed a modular approach (to divide the content and the interfaces) to help to organise the workflow, namely, density inference, space-time interpolation and event geo-location (spatial mapping). Finally, they argue for accessing the networks' data without having them in-house, but by bringing the algorithm to the data.

The next session included two invited presentations on Big Data and were followed by a large external audience.Antonio Argüeso (National Institute of Statistics – INE, Spain) gave insights on the use of Big Data within the 2021 census of Spain. Even if this new census built upon administrative registers will bring many advantages, such as quality (better measurement of reality), timeliness, continuity or more information (not limited by a census form), it will also present important limitations, e.g. some of the information cannot be retrieved from administrative registers or, even if found, it can be biased. The use of mobile phone data, through the Mobile Network Operators (MNO), could solve part of the challenges. However, it has been difficult until now to retrieve these data due to the concerns from MNO about the access to individual data or the enormous importance that Big Data has nowadays for companies (which are offered at market prices). As a solution until there is more consensus on which data to use, Argüeso proposed combining traditional methods and private sources. On the other side, Alvaro Ortiz (BBVA Research) provided the experience of a private enterprise using Big Data to monitor world geopolitics. He presented the exploitation of the Global Database on Events Location and Tone (an open database with georeferenced events with more than 3000 themes and emotions), through text mining and sentiment analyses to detect social unrest events, dynamic migration flows or global health issues. Moreover, he showed the results of a comparison between aggregated and anonymised BBVA Big Databases on card transactions and the INE statistics in Spain on Retail Trade Indices, with a correlation of 95%. In Mexico, this has been used to observe economic recovery time for natural disasters or to monitor tourism, which is very useful for governments and tourism management.

The closing session of the workshop was dedicated to two papers on Twitter data. First, Dariya Ordanovich (ESRI España) offered an overview of the interdisciplinary work carried out with her colleagues on using geotagged messages from Twitter for fertility nowcasting, which introduces a significant added value to the statistical production at marginal cost. The intention was to understand the fertility intentions and the short-term fertility changes in time and space.

After processing the language, the next step was to build a machine learning model, by filtering and classifying tweets (2.500 of them, manually), and then testing text classification algorithms. Findings revealed that geocoded Twitter data might serve as a dynamic agent for detecting the changes in the overall fertility patterns. José Javier Ramasco (Institute for Cross-Disciplinary Physics and Complex Systems, Spain) and colleagues, in collaboration with UNICEF, used geocoded Twitter data to detect migration flows, focusing in Venezuela. They explored questions on travelling routes, exit times, spatial distribution on new settlement areas, etc. At a technical level, Ramasco also explained how to filter the data and to eliminate bots among the Twitter users (e.g. if the user appears to be moving faster than a plane), as well as to extract information from the tweets. In conclusion, their results showed that these new digital data can be an important complement to surveys, police stations and border control.

The workshop was closed with a summary and discussion by Emilio Zagheni, Francesco Billari and Diego Ramiro, as well as a proposal to pursue this work within the IUSSP Panel on Digital Demography's programme of activities in the next months.