Big Data and alternative data sources on migration:

From case-studies to policy support

European Commission – Joint Research Centre (JRC), Ispra (Italy)

30 November 2017

Summary report

Marzia Rango[1] and Michele Vespe[2]

Over the past few years, migration has risen as one of the most challenging issues confronting policymakers around the world. This has been particularly true for countries in the Global South grappling with displacement caused by natural disasters, violence and human rights abuses, but also for countries in the North – namely in Europe – which saw relatively large increases in inflows of asylum-seekers from poor, unstable or war-torn countries. The growing complexity of internal and cross-border human mobility has highlighted the need for reliable and timely data to inform humanitarian assistance and policy responses – a need that traditional statistical systems are often not well-equipped to meet.

Meanwhile, most of the data that exists in the world today are not collected by national statistical offices but by private companies or international agencies. Estimates suggest that about 90% of world's data were generated only in the last two years. Technological innovations and the reduction in the cost of digital devices have meant that vast amounts of data generated through use of mobile phones, internet-based platforms and other digital devices are now collected in real time, at very little cost. Sensible and responsible analyses of these "digital crumbs" have the potential to offer important insights into societal phenomena, including migration, as demonstrated by a number of promising applications. However, a series of issues, ranging from access and analytical difficulties to privacy and security risks, mean that such vast potential still remains largely untapped.

This workshop[3] gathered representatives from the academic, policy and business communities to discuss the state-of-the-art in data innovations in the field of migration, and identify possible ways to bring analytical insights from new data sources closer to migration policy needs.

[1] International Organization for Migration, Global Migration Data Analysis Centre (GMDAC).
[2] European Commission, Joint Research Centre.
[3] See https://bluehub.jrc.ec.europa.eu/bigdata4migration.

Background

Despite growing efforts from national governments and the international community, data on migration are still commonly agreed to have serious limitations. National population censuses – traditionally the main source of data on migrant stocks in most countries – are infrequent or not conducted regularly, and cannot therefore provide timely information on migration dynamics. Migrants, particularly those on irregular status, are often absent from sampling frames used for household surveys; if they are not able to access services in their host country, they will not appear in administrative records either. These limitations, particularly in countries with large migrant populations, create serious gaps in the information needed to design effective policies and to monitor the needs of the population and how these change over time. If migrants are absent from the data, they will be absent from policy and planning processes, with negative consequences for their well-being and their ability to positively contribute to development in host and home countries alike.

The long-standing calls to improve data on migrants and migration have received great momentum in recent years. The historic inclusion of migration-related targets in the Sustainable Development Goals revealed weaknesses in the data needed to both implement and monitor the far-reaching commitments made by the international community. The historic adoption of a Global Compact for Safe, Orderly and Regular Migration in December 2018 will also place further demands for data upon states and the international community in the follow-up and review phase. Calls for a *data revolution* resulted in the establishment of the UN Secretary General's Independent Expert Advisory Group on the Data Revolution – a group of experts from both public and private sectors – whose report, *A World That Counts*, included a series of recommendations to fill data gaps and harness emerging technologies to improve data and reduce inequalities in access to information.

A series of other initiatives and bodies were born at the UN-level. The recognition that the potential of big data to improve information on a number of development-related areas led to the decision of the UN Statistical Commission to create a Global Working Group (GWG) on Big Data for Official Statistics in 2014.[4] The GWG, whose members include the European Commission (Eurostat), OECD and the World Bank, is tasked with investigating 'the benefits and challenges of Big Data,' including by identifying concrete examples of the potential of using these new data sources for monitoring SDG indicators, and addressing issues of quality, confidentiality and feasibility of using big data. The same recognition of the opportunity offered by digital data brought to the creation by the UN Secretary-General of the UN Global Pulse, a network of 'innovation labs' conducting research on big data for development, in partnership with experts from a range of sectors.[5] Also, in 2015 the UN Data Innovation Lab was established by the UN Chief Executives Board for Coordination to 'enable the UN to harness the power of Data Revolution for Sustainable Development.'[6] This initiative consists in a series of thematic workshops led by different UN Agencies to share knowledge on data innovation projects.

[4] See https://unstats.un.org/bigdata/.
[5] See https://www.unglobalpulse.org/about-new.
[6] See https://data-innovation.unsystem.org/background-data-innovation-lab.

At the European level, a number of initiatives recently originated to better understand the potential impact of big data to provide insights during the entire policy cycle, from policy planning and design to implementation and evaluation. In 2016, as part of a broad Big Data Action Plan and Roadmap, the European Statistical System (ESS)[7] launched the ESSnet Big Data project[8] to integrate big data in the regular production of official statistics. In addition, Eurostat established a dedicated task force to explore the potential of big data for producing European statistics and informing EU policies.

The focus on data innovation, not only at the international but also national levels, has brought together statisticians, data scientists, civil society groups, policymakers and private sector representatives with a shared interest in exploring how new technologies and innovations can be combined with existing methods to improve the coverage, timeliness and quality of data, and to ensure that it is used to solve problems and to monitor progress. Experiments in bringing together old and new sources and methods are underway in a number of sectors – including human mobility – to establish new ways of collecting and analyzing data.

However, some skepticism and uncertainty among statisticians and policymakers still remains on the feasibility of using big data sources to address gaps in migration statistics and inform policymaking. A global survey conducted by the UN GWG on Big Data on big data management and existing big data projects revealed that less than a third of the 32 OECD countries agreed that big data would be useful to meeting new demands for data such as for SDG monitoring; responses were more positive among the 61 non-OECD countries who completed the survey, with 67% agreeing on the potential of big data to meet such new demands.[9] The UN GWG also hosts an inventory of big data projects that are relevant for official statistics: only a small number of the projects listed focus on mobility or demographic and social statistics, and most of them regard internal mobility.[10]

Importantly, there is currently *no dedicated unit tasked with investigating the potential of big data and innovative data sources for analysis of migration-related trends and aspects*. Global and national initiatives on the topic seem scattered, and the new bodies or mechanisms created to harness the data revolution for sustainable development do not specifically focus on realizing the potential of big data for measurement of *international migration*.

The workshop

The workshop *Big Data and alternative data sources on migration: from case studies to policy support* – jointly organized by the European Commission's Knowledge Centre on Migration and Demography

[7] ESS is a partnership between Eurostat and national statistical institutes or other national authorities in each European Union (EU) Member State responsible for developing, producing and disseminating European statistics. See http://ec.europa.eu/eurostat/web/ess/

[8] See https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data

[9] Report of the Global Working Group on Big Data for Official Statistics, E/CN.3/2016/6, Statistical Commission, 47th session, 8—11 March 2016 (p.4).

[10] See https://unstats.un.org/bigdata/inventory/.

(KCMD) and IOM's Global Migration Data Analysis Centre (GMDAC) brought together researchers, private sector representatives, policymakers and practitioners to:

i. Review the latest examples of concrete applications of big data and alternative data sources in the field of migration;
ii. Identify the most promising applications of big data and alternative sources to complement and enrich traditional data on migration;
iii. Discuss possibilities, obstacles, and requirements for more systematic use of big data in migration, and for increased collaboration between researchers, data providers and policymakers;
iv. Suggest pragmatic follow-up steps to expand use of such data sources in support of migration policymaking.

The workshop was structured into **three sessions**.

Session I – Recent applications of big data and alternative data sources to the study of migration-related trends

Session I aimed to explore the potential of big data for the analysis of migration-related patterns based on existing studies and projects. Participating researchers presented selected applications of big data and innovative data sources with relevance to migration. These included using data from the Facebook advertising platform as "real-time census data" as presented by Ingmar Weber, Research Director at the Qatar Computing Research Institute (joint work with Emilio Zagheni, University of Washington). For instance, based on Facebook data, there would be approximately 214 million "expats" in the world – people stating to be living in a country other than their self-reported "home country" – which is not far from the 2017 estimate of the total number of international migrants globally of 258 million.[11] There are clearly issues with using Facebook data to estimate (international) migrant stocks, particularly selection bias (the population of Facebook users may not be representative of the population at large), reliability of self-reported information (subjectivity over what the user may consider as his or her "home country," for instance), and the difficulty in applying the UN-recommended definition of an international migrant when using Facebook data, given the frequent lack of information on length of residence in the host country. Scholars are however working on reducing selection bias via model fitting and results are promising. Also, estimates of the foreign-born population using Facebook data may be particularly helpful in countries that lack recent census data and that show high Facebook penetration rates.

The combination of independent data sources with different spatio-temporal resolution and attributes was highlighted by more than one speaker as key to improving understanding of certain phenomena. For instance, mobile phone records of calls made from a specific city to a foreign country, coupled with

[11] UN Department of Economic and Social Affairs (2017), see https://www.un.org/development/desa/publications/international-migration-report-2017.html.

traditional census statistics, could contribute to analyzing patterns of migrant integration, as well as residential and mobility segregation, as demonstrated by Fabrizio Natale of the European Commission Joint Research Centre. Use of the social media platform Instagram among refugees stranded in Greece, together with qualitative information about the accounts, could help as an early-warning system for refugee movements, as shown by Stefano Iacus, Professor of Statistics at the University of Milan. In this sense, social media data may be more helpful to identify or anticipate trends in people's movements, rather than quantifying these movements. Technological sustainability of services based on big data and alternative data sources was also highlighted as a key issue: the availability of data can change because of new data access restrictions or due to discontinuity of the source.

Mobile phone call records combined with satellite data can be used to build "dynamic population maps," as shown by Alessandro Sorichetta, researcher at the University of Southampton affiliated with the Flowminder foundation. Sorichetta also presented his analysis of call detail records and cross-border flow estimates in southern Africa to identify movements between cross-border communities, which could help track the spread of malaria.

Mobile phone call detail records (CDR) can also be used to identify the so-called "transnationals", or people living or working in more than one location, as presented by Rein Ahas, Professor at the University of Tartu, Estonia. Transnationals could be defined, according to Ahas, as people "keeping the SIM card of their home country but living and working abroad." Transnationalism patterns are difficult to capture through traditional data sources, though there are of course methodological issues with CDR data, namely of validation. However, while CDR data are generally more helpful to analyse internal migration patterns, they could potentially also be used to study international migration across neighbouring countries, for instance, thus possibly complementing flow data from administrative sources.

Social media content, for instance from Twitter, can also be used to analyse how opinions – including on migration – can become polarized and self-referential, how separate "echo-chambers" of opinions are formed and network clustering occurs, as shown by Natale. Natale also presented a project involving daily monitoring of thousands of news websites in different languages to identify the source of fake news and how they propagate. According to Dino Pedreschi, Professor at the Università di Pisa in Italy, big data sources could be used to build a "super-diversity" index, moving beyond residential segregation, as well as an "integration index" based on, among others, food-purchasing behavior of foreigners in a certain locality.

Big data for migration: summary of challenges and opportunities:

Challenges	Opportunities
– *Data access (proprietary data, technological sustainability, costs);* – *Noisy data, analytics to process and filter;* – *Pilot studies, no systematic services;* – *Confidentiality, security, and ethical concerns;* – *Sampling bias;* – *Difficulty in applying UN-recommended definition of international migrant.*	– *High spatial resolution;* – *High frequency of update;* – *Timeliness, virtually real-time;* – *Global coverage, even in areas with limited/no migration statistics;* – *Not bound to statistical definitions: potential to better understand temporary forms of migration;* – *Larger sample sizes compared to surveys.*

Finally, use of predictive analytics can be helpful for disaster displacement preparedness and response, and early warning of human rights violations, conflict and xenophobia, as presented by Rebeca Moreno Jimenez of UNHCR Innovation Service.

During his discussion of presentations made in Session I, Jean-Christophe Dumont of OECD emphasized the need to consider whether current big data efforts are matching migration measurement objectives, and think of what these objectives are, including by identifying the main migration knowledge gaps to be filled. He also raised the need to have flexible definitions of migration, given the difficulty in tracking length of stay of the foreign-born or foreign citizens via big data sources, and that of measuring the margin of error of big data analyses of human mobility, so to facilitate use of such analyses for policymaking purposes – and assess the risks in doing so.

Session II – Data holders/providers: perspectives and applications

Session II aimed to provide examples of how data providers can contribute to more systematic analyses of human mobility through big data, not only by sharing raw data, but by providing services and devising applications that can help measure migration-related patterns. Zbigniew Smoreda, Sociologist at Orange Lab presented the Open Algorithms project (OPAL), a collaborative project aimed to harness the power of big data and advanced analytics for the public good, launched in 2016. OPAL provides a model for how to share data while ensuring the protection of privacy and analytical flexibility, through legally and technically certified algorithms. The project is currently being piloted in Colombia and Senegal, with the aim of developing a Beta platform and open-source algorithms. Frederic Pivetta, Co-founder and Managing Partner at Dalberg Data Insights, presented the work of his team, which acts as an intermediary between data providers and data users to achieve social impact in a variety of countries and sectors – from financial inclusion to urban mobility, health and agriculture. The idea of Dalberg Data Insights is to "bring the algorithm to the data," meaning that private data can be accessed without it needing to leave the premises of the data provider (e.g. the mobile phone operator), so that customers' privacy is not compromised and proprietary data is not revealed. Similarly to the OPAL project, data

here is seen as a service, so Dalberg Data Insights focuses on creating specialized analytics that can be used by governmental and non-governmental entities in tackling development challenges.

Fernando Reis, Big Data Statistician at Eurostat, presented the activities of Eurostat's Task Force on Big Data, which currently has several pilot projects in various areas (though not yet on migration). One of the aims of the Task Force is to establish a dialogue between National Statistical Offices (NSO) and big data holders, such as Mobile Network Operators (MNO), to produce "trusted smart statistics" that can be used for policymaking. The collaboration between NSOs and MNOs is particularly important given that the former have a legal mandate to produce statistics, a legal basis to collect data as well as corresponding obligations to ensure confidentiality, and the latter hold large amounts of data collected in real time and potentially available at a lower cost than with traditional sources.

Finally, Mirek Pospisil, EU Public Policy and Government Affairs Manager at LinkedIn, provided examples of possible analyses of the data the social network holds on its more than 530 million members globally. The data allows, for instance, for a digital mapping of the workforce in certain regions or countries, including of migrants and their skills. It can contribute to track movements of highly skilled migrants through self-reported changes in their job positions, particularly in locations and sectors where penetration rates of LinkedIn are relatively high.

A discussion among the workshop participants highlighted some of the issues with private sector engagement related to trust and reputation, lack of control over potential results and consequences of data sharing, or reluctance towards "data commercialization". This may be due to fear of potential negative consequences while regulation in this innovative area is still shaky, and possible internal tensions within private companies, usually between the research and policy departments, with a resulting need to justify internally why data sharing for public purposes should be pursued.

Piotr Juchno, statistician at Eurostat, provided a summary of lessons learned and mentioned some of the challenges related to private-public partnerships for analyses of social phenomena, such as data protection issues and feasibility of use of big data for public policy choices. Juchno particularly emphasized the need to connect the big data scientist and statistician communities, in view of the skepticism and unfamiliarity statisticians may have with innovative data sources, but also to improve awareness in the big data scientific community of the context, requirements and definitions that official statistics of migration entail.

Session III – Roundtable discussion

The final session aimed to discuss the main challenges related to uses of big data for policymaking – i.e. issues of privacy, data access, as well as technical, analytical and legal challenges – provide recommendations to help overcoming such barriers, and suggest possible strategies to leverage (big) data for policymaking.

The session opened with a keynote speech by Stefaan Verhulst, Co-Founder and Chief Research and Development Officer of the Governance Laboratory (GovLab) at New York University. The presentation

stressed the need to rethink private-public partnerships in order to enable use of private data for public good. Verhulst described the various ways in which private data can inform public policy and explained different models of so-called "data collaboratives," all of them depending on the trade-off between the openness of data and the need for collaborations to gain access to the data. One of these models of data collaboratives involve the creation of intelligence products within the private sector, that are able to generate insights from the data. Such insights can be shared by the private sector without necessarily sharing the raw data they are based on. Other models are more heavily reliant on data sharing on the part of the private sector (see the table below with concrete examples presented during the workshop).

Data "Collaboratives" and concrete examples from the workshop

Data "Collaboratives"	**Examples**
Data Cooperatives or pooling	Mobile Network Operators partnerships - Eurostat[12]
Intelligence products	LinkedIn Economic Graph[13]
Prizes & Challenges	Data 4 Development[14] – Data 4 Integration[15]
Application Programming Interfaces	Facebook / LinkedIn advertising platform[16]
Research partnerships	JRC maps of local communities using micro-census data[15]
Trusted intermediary	Dalberg Data Insights[17]

Companies have certain incentives to indeed share data, such as, for instance, reputation and retention of talent (e.g. data scientists), access to analytical insights, generation of revenue (if data is shared for profit), but also regulatory compliance and corporate social responsibility. They may, however, show concerns in relation to privacy and security of their customers, representativeness and quality of the data, and a company culture discouraging data sharing. A new framework to share data responsibly is therefore needed, also accounting for the opportunity cost of not sharing the data. In order to build the foundation to overcome the large transaction costs of creating data collaboratives, some steps would be necessary, such as the professionalization of "data stewards" – figures in charge of identifying relevant data sources and ways to access them; the identification, on the part of policymakers, of questions to be answered; the systematic collection of evidence of what works and what does not; and the creation of a "movement" advocating for open data for the public good. According to Verhulst, a key question to address would be what kind of data collaborative would be needed for a certain policy proposition –

[12] Fernando Reis, *The use of big data in official statistics*, "Big Data and Alternative Data Sources on Migration: from Case Studies to Policy Support", 30 November 2017, Ispra.
[13] Mirek Pospisil, *LinkedIn Economic Graph: a focus on Europe*, "Big Data and Alternative Data Sources on Migration: from Case Studies to Policy Support", 30 November 2017, Ispra.
[14] Zbigniew Smoreda, *Mobile phone data for migration analysis: new possibilities, risks and difficulties*, "Big Data and Alternative Data Sources on Migration: from Case Studies to Policy Support", 30 November 2017, Ispra.
[15] Fabrizio Natale, *Big data and migration, Buzz-phrase or policy relevant applications*?, "Big Data and Alternative Data Sources on Migration: from Case Studies to Policy Support", 30 November 2017, Ispra.
[16] Ingmar Weber, *Using Internet Advertising Data for Studying International Migration*, "Big Data and Alternative Data Sources on Migration: from Case Studies to Policy Support", 30 November 2017, Ispra.
[17] Frederic Pivetta, *Big Data, Social Impact and concrete examples on Migration,* "Big Data and Alternative Data Sources on Migration: from Case Studies to Policy Support", 30 November 2017, Ispra.

which places policy needs at the core of finding sustainable solutions for big data analyses in any field, including migration.

During the discussion that followed, one of the policymakers participating in the workshop mentioned some of the key knowledge gaps in migration – including regular migration flows and migrants' characteristics, migrant smuggling trends and migration forecasting – stating that having data characterized by a large margin of error would still be better than having no data at all. The need to talk about data innovation and policy needs to policymakers at all levels, including in particular those at the local level, was also raised during the discussion.

Conclusions and way forward

The workshop provided a timely and unique opportunity to take stock of the vast potential of big data and innovative data sources for analysis of migration-related aspects through an active dialogue among policymakers, scientists and the private sector. What clearly emerged is that initiatives in this area are numerous and rapidly expanding, though it may at times be difficult a) to identify how new data sources can complement traditional data sources without compromising agreed definitions of migration, and b) to assess the degree to which information generated by such sources can be effectively used by policymakers.

The workshop demonstrated the opportunities offered by new data sources, such as the timely availability of data and analysis, the wide coverage – especially in the case of mobile phone users – and the relatively low costs at which data can be collected. In addition, concrete examples showed the potential to contribute to the understanding of phenomena that are currently hard to capture through traditional data sources, such as migrant return, inclusion and integration, migration flows, particularly within the Global South, intra-regional mobility and migration (even of EU citizens within the EU), circular and seasonal mobility patterns, irregular migration and migrant smuggling, human trafficking and early warning of displacement.

These opportunities are mirrored by numerous challenges. First, there are ethical and privacy issues deriving from using data automatically generated by users, as well as civil liberties concerns due to the risks of using such data for surveillance purposes. Second, big data raises issues of a potential widening of the "digital divide" between information-rich and information-poor countries, and risks of an exacerbation of global inequalities. Third, there are technical and analytical challenges in using big data sources, due to difficulties in accessing data – mostly held by private actors – inappropriate infrastructure and data management and security systems, methodological difficulties in extracting meaning from huge and complex volumes of data, and the need to address the inherent self-selection bias, as big data derives from users of mobile and internet-based platforms and are therefore not necessarily representative of the population at large. However, while these challenges may still hamper use of big data sources for policymaking, the workshop shed light on possible strategies to address some

of them, by suggesting new frameworks to enhance collaborations between researchers and data scientists, the private and the public sector.

To overcome the challenges and systematize use of big data sources for migration research and policymaking, investments seem to be particularly needed in the following areas (which are not unique to migration):

1) The establishment of an **adequate regulatory and legislative framework** for the collection, analysis, and sharing of big data; an **international dialogue** between regulators, big data users and providers should be the starting point.

2) The **upgrading of infrastructure and security systems** at the **national level**.

3) The creation of incentives towards the **development of private-public partnerships** for data exchange and collaborations.

4) **More research** on ways to capitalize on innovative data sources in the field of migration, and systematic ways to take stock of existing applications and collate existing knowledge.

As a way to facilitate these investments, the European Commission's Knowledge Centre on Migration and Demography (KCMD) and IOM's Global Migration Data Analysis Centre (GMDAC) are planning to create a **Big Data for Migration Alliance (BD4M)** to advance discussions on how to harness the potential of big data sources for the analysis of migration and its relevance for policymaking, while ensuring the ethical use of data and the protection of individuals' privacy.

As conveners of BD4M, the KCMD and IOM's GMDAC would welcome the participation of representatives from international organizations and non-governmental organizations, members of national statistical offices, private sector representatives, researchers and data scientists interested in contributing, in various capacities, to realizing the potential of big data to complement traditional data sources on migration.

The BD4M would be global in scope, and would aim to make progress in the area of big data and migration through **3 main areas of work**:

- **Awareness-raising and knowledge-sharing**: this would involve taking stock of new developments and applications of big data sources in the field of migration, and promoting the sharing of knowledge, including through dedicated publications and conferences. It could also involve the mapping of existing big data and migration initiatives at the national and EU/UN level and the creation of a regularly-updated repository of big data and migration projects, to identify good practices and possible synergies.
- **Capacity-building**: this would entail providing support to countries interested in identifying ways to make use of big data sources to complement official statistics on migration, including through specific training modules as well as need-based technical assistance and guidance, in collaboration with relevant partners and agencies.

- **Policy-oriented research**: members of the Alliance would develop research projects aimed at testing new applications of big data in the field of migration, as well as addressing the technical, methodological and ethical challenges associated with uses of big data and other innovative sources for research purposes. This will include the identification of a number of 'priority areas', based on major gaps in traditional migration statistics and regular consultations with policymakers, to translate policymakers' needs into data requirements.

The BD4M would also encourage the creation of a **network of "data stewards"** across private and public institutions, to facilitate exchanges of information and good practices on how to leverage big data for migration analysis, with due consideration of privacy and ethical concerns.