Received: 2 February 2022 | Accepted: 3 February 2022

DOI: 10.1111/imig.12984

COMMENTARY

INTERNATIONAL MIGRATION IOM UN MIGRATION

Can big data deliver its promises in migration research?

Albert Ali Salah[1,2]

[1]Information and Computing Sciences, Utrecht University, Utrecht, the Netherlands

[2]Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey

Correspondence

Albert Ali Salah, Buys Ballot Gebouw, Utrecht University, Princetonplein 5, 3584CC Utrecht, the Netherlands.
Email: a.a.salah@uu.nl

Funding information

This article was funded by H2020 Research and Innovation Programme, Grant/Award Number: 870661.

As a computer scientist working on algorithms and applications for human behaviour analysis since the 90s, I was able to witness the rise of "big data," which was jokingly defined by one of my university professors as "data that would not fit on your computer". Indeed, such is the volume of data constantly generated by humans moving about with mobile phones in their pockets, using devices and services that create, store and share information about such usage, contributing and propagating content on social media, that it becomes difficult to organise, classify and interpret these digital traces with traditional computational approaches. Add to this the devices and systems that observe humans, such as surveillance cameras, smart infrastructure elements that log users, remote sensing and satellite systems, and the picture becomes even more complex. Yet, with ample opportunities for monetisation, approaches and systems were quickly developed to process big data. In particular, companies quickly recognised that big data collected from customers could provide valuable insights for marketing and optimisation tasks, for content customisation, user adaptation and modelling. But such data were also valuable to governments for decision-making and policy, and for many researchers interested in human behaviour, including those in the field of migration research, looking for reliable and granular indicators of human movements over the globe.

The main premise of *computational social science* (or *social computing*) is that large-scale and complex human behavioural data, typically stored by companies that benefit from it for the mentioned purposes, can be analysed and re-purposed to address research questions from the social sciences (Lazer et al., 2020). In this new paradigm, carefully constructed samples and in-depth questions were replaced by numbers of subjects greater by orders of magnitude, and simple indicators that were sampled across much greater temporal and spatial resolutions than the traditional "snapshots". In addition, these were analysed in an aggregated fashion to provide proxy variables. For example, in a seminal study related to migration research, Blumenstock et al. (2015) illustrated that a person's mobile phone usage history could be used to create a wealth indicator. They used mobile phone call detail records (CDR) collected for accounting purposes by the phone company and containing the base tower locations, times, and durations of phone calls for each customer. Aggregating anonymised data from 1.5 million customers, Blumenstock and co-authors created a detailed map that predicted the wealth of people living in Rwanda, which showed a high correlation with the last two Demographic and Health Surveys conducted with seven and thirteen thousand people, respectively, at a fraction of the cost, and in a very timely fashion. Does this mean that similar big data initiatives can be easily created all over the world to enable improved analyses of factors known to relate to migration (Sîrbu et al., 2021), or to provide timely insights to policy makers? Hardly so; there are technical, ethical, legal, and as Scheel and Ustek-Spilda (2018) highlighted, political hurdles that hinder adopting big data solutions in migration statistics.

Before continuing further, I would like to point out that the big data usage examples I advocate work with anonymised and aggregated data, and do not—in principle—involve data, which can be used to identify specific individuals. In the last few years, several instances of technology usage in areas related to migration management came into question (Molnar, 2021). Particularly, high-tech solutions for large-scale biometric identification of refugees, drone and robotic surveillance in the borders, automatic decision-making based on artificial intelligence solutions in sensitive areas where the cost of an error cannot (and should not) be quantified, including lie detection technologies at European borders have been discussed in terms of their ethical and human rights ramifications. Molnar (2021, p.134) also pointed out that Covid-19 increased the investment in technological solutions and states that "as governments move towards biosurveillance to contain the spread of the pandemic, there has been an increase in the use of tracking, automatic drones and other types of technologies that purport to help manage migration, exacerbating potential human rights concerns (Cliffe, 2020; Lewis & Mok, 2020; Molnar & Naranjo, 2020)." These are all valid concerns, but it is possible to process big data with proper checks and balances, controlling for both individual and group privacy (Salah, Canca, et al., 2022), and we should weigh their usefulness in the context of specific risks. After all, flexible and ubiquitous technologies such as mobile phones do have many uses - such as in helping to contain the spread of a pandemic (Oliver et al., 2020), and the existence of technology—and associated data—brings an obligation with it, in that if it is possible to responsibly harness it for improving the lives of people, it should be considered seriously by governmental and non-governmental actors (Letouzé & Oliver, 2019).

The primary potential of big data for migration research seems to be in addressing *data gaps* (Bircan et al., 2020; Bosco et al., 2022), which are created by, for instance, inconsistencies in the definitions and data collection methodology, lack of adequate statistics and plainly lacking data on irregular migration. Different data types may provide information about such gaps (Salah, Bircan, et al., 2022). Mobile phone call detail records contain very detailed mobility information, but some exceptions aside, are processed without linking it to demographic information, to protect the privacy of the data subjects. Nonetheless, it can be used to produce wealth indicators, show the concentration and mobilities of groups, indicate infrastructure usage, and even provide indicators of social integration by modelling intergroup encounters and communication (Bakker et al., 2019). Remote sensing provides population estimates that are considered valuable information in contexts of natural and man-made disaster areas and climate-related mobility (Bircan, 2022). Social media data can be a particularly rich source offering different insights depending on the platform. For example, Twitter can provide information about sentiments of and about migrants and refugees while LinkedIn can offer indicators about skilled migration, and Facebook on demographics (Bosco et al., 2022; Coimbra et al., 2022; Kim et al., 2022). Combining multiple data sources can potentially be even more powerful; for example, refugee mobility and settlement can be observed via mobile phone data, and real-estate price averages can be linked to home locations, so examining these two sets of data together provides an opportunity to estimate refugee wealth in greater detail (Bertoli et al., 2019).

What are the premises under which data gaps are filled? First of all, data access is an important issue. Some of these data sources are publicly available for research (e.g. remote sensing data with limited resolution), but others are buried in the servers of private companies and very difficult to access (e.g. mobile CDR). Second, processing such high volumes of data requires mastery of data scientific tools, possibly including database management systems, scripting tools and analysis tools that range from natural language processing to image processing, depending on the modality. Ideally, this calls for interdisciplinary collaborations between computer scientists and migration scholars. The third premise is the translation of migration scholars' inquiries (or the data gap) into the disciplinary language of computer scientists while simultaneously avoiding reductionisms typical for this field. For instance, if social integration is to be assessed via mobile phone data, quantitative indicators should be found that are of relevance to social integration, and they need to be validated with some approach. Blumenstock et al. (2015) used previous census data to validate their method, but if there is a data gap to be addressed, there may not be such an obvious choice. The assumptions under which the approximations and projections would work should also be determined in collaboration to account for not only domain expertise, but also biases that may come from the algorithmic models employed. Finally, applied ethics expertise is required to properly assess the potential risks, and this topic is mostly lacking from most technical computer science curricula. The ethical review is necessary not only for the study itself but also for the communication of results, given the sensitivity that surrounds how the findings are communicated.

As a way of addressing the first difficulty, Verhulst and Young (2019) proposed *data collaboratives*, which are public-private partnerships based on data owned by private parties. Different models have been proposed for such collaborations (Letouzé & Oliver, 2019), but the main idea is that a private dataset is re-purposed for public good. *Data challenges* are a form of data collaboratives, where the private data are processed to remove personal information and opened to a larger research community, allowing multiple groups to analyse it from different aspects (Salah, 2021). An example of such challenges was the Data for Refugees (D4R) project supported by TUBITAK, UNHCR, UNICEF and IOM, where Türk Telekom opened a mobile CDR dataset collected in Turkey from 1 million users over a year to provide insights into Syrian refugee mobility in Turkey with the aim of improving their living conditions (Salah et al., 2019). This challenge, with 60+ participating research teams (and an ethics committee examining both the projects and resulting publications) was useful for capacity building, initialising inter-disciplinary and international collaborations, and creating some policy recommendations. However, a challenge is (typically) a one-off event and sustainable data processing requires longer relationships, where the infrastructure must be created to share computed indicators instead of the data itself. Furthermore, to be truly useful, data-driven recommendations must be taken up by researchers with deep domain expertise and by policy makers, further investigated, combined and triangulated with existing information sources and evaluated by taking into account other qualitative factors influencing research and policy considerations. Bridging the data gaps requires actions from all involved parties, including private data holders, policy makers, and most importantly, migration scholars and data scientists bridging disciplinary language gaps.

DISCLAIMER

The opinions expressed in this Commentary are those of the author and do not necessarily reflect the views of the Editors, Editorial Board, International Organization for Migration nor John Wiley & Sons.

ORCID

Albert Ali Salah https://orcid.org/0000-0001-6342-428X

REFERENCES

Bakker, M.A., Piracha, D.A., Lu, P.J., Bejgo, K., Bahrami, M., Leng, Y. et al. (2019) Measuring fine-grained multidimensional integration using mobile phone metadata: the case of Syrian refugees in Turkey. In: Salah, A.A., Pentland, A., Lepri, B. & Letouzé, E. (Eds.) *Guide to mobile data analytics in refugee scenarios*. Cham: Springer, pp. 123–140.

Bertoli, S., Cintia, P., Giannotti, F., Madinier, E., Ozden, C., Packard, M. et al. (2019) Integration of Syrian refugees: insights from D4R, media events and housing market data. In: Salah, A.A., Pentland, A., Lepri, B. & Letouzé, E. (Eds.) *Guide to mobile data analytics in refugee scenarios*. Cham: Springer, pp. 179–199.

Bircan, T. (2022) Remote sensing data for migration research. In: Salah, A.A., Korkmaz, E.E. & Bircan, T. (Eds.), *Data science for migration and mobility*. Proceedings of the British Academy. London: British Academy / Oxford University Press.

Bircan, T., Purkayastha, D., Ahmad-Yar, A.W., Lotter, K., Iakono, C.D., Göler, D. et al. (2020) *Gaps in migration research. Review of migration theories and the quality and compatibility of migration data on the national and international level, Deliverable 2.1 of the HumMingBird project*. Available from: https://hummingbird-h2020.eu/images/publicationpdf/d2-1-eind-1.pdf [Accessed on 29 January 2022]

Blumenstock, J., Cadamuro, G. & On, R. (2015) Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264), 1073–1076.

Bosco, C., Grubanov-Boskovic, S., Iacus, S., Minora, U., Sermi, F. & Spyratos, S. (2022) Data innovation in demography, migration and human mobility, EUR 30907 EN, Publications Office of the European Union, Luxembourg, 2022, ISBN 978-92-76-46702-1, JRC127369.

Cliffe, J. (2020) *The rise of the bio-surveillance state*. Available from: https://www.newstatesman.com/uncategorized/2020/03/rise-bio-surveillance-state [Accessed on 29 January 2022]

Coimbra, C., Fatehkia, M., Garimella, K., Weber, I. & Zagheni, E. (2022) Using Facebook and LinkedIn data to study international mobility. In: Salah, A.A., Korkmaz, E.E. & Bircan, T. (Eds.) *Data science for migration and mobility*. Proceedings of the British Academy. London: British Academy / Oxford University Press.

Kim, J., Pollacci, L., Rossetti, G., Sîrbu, A., Giannotti, F. & Pedreschi, G. (2022) Twitter data for migration studies. In: Salah, A.A., Korkmaz, E.E. & Bircan, T. (Eds.) *Data science for migration and mobility*. Proceedings of the British Academy. London: British Academy / Oxford University Press.

Lazer, D.M.J., Pentland, A., Watts, D.J., Aral, S., Athey, S., Contractor, N. et al. (2020) Computational social science: Obstacles and opportunities. *Science*, 369(6507), 1060–1062.

Letouzé, E. & Oliver, N. (2019) *Sharing is caring: four key requirements for sustainable private data sharing and use for public good*. London: Data-pop Alliance and Vodafone Institute for Society and Communications.

Lewis, S. & Mok, O. (2020) *Malaysia enforces lockdown compliance with drones*, *Privacy International*. Available from: https://privacyinternational.org/examples/3509/malaysia-enforces-lockdown-compliance-drones [Accessed on 29 January 2022]

Molnar, P. (2021) Robots and refugees: the human rights impacts of artificial intelligence and automated decision-making in migration. In: McAuliffe, M. (Ed.) *Research handbook on international migration and digital technology*. Cheltenham, UK: Edward Elgar Ltd, pp. 134-151.

Molnar, P. & Naranjo, D. (2020) *Surveillance won't stop the coronavirus*. Available from: https://www.nytimes.com/2020/04/15/opinion/coronavirus-surveillance-privacy-rights.html [Accessed on 29 January 2022]

Oliver, N., Lepri, B., Sterly, H., Lambiotte, R., Deletaille, S., De Nadai, M. et al. (2020) Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle, *Science*. *Advances*, 6(23), eabc0764.

Salah, A.A. (2021) Mobile data challenges for human mobility analysis and humanitarian response. In: McAuliffe, M. (Ed.) *Research handbook on international migration and digital technology*. Cheltenham, UK: Edward Elgar Ltd, pp. 107-122.

Salah, A.A., Bircan, T. & Korkmaz, E.E. (2022) New data sources and computational approaches on migration and human mobility. In: Salah, A.A., Korkmaz, E.E. & Bircan, T. (Eds.) *Data science for migration and mobility*. Proceedings of the British Academy. London: British Academy / Oxford University Press.

Salah, A.A., Canca, C. & Erman, B. (2022) Ethical and legal concerns on data science for large scale human mobility. In: Salah, A.A., Korkmaz, E.E. & Bircan, T. (Eds.) *Data science for migration and mobility*. Proceedings of the British Academy. London: British Academy / Oxford University Press.

Salah, A.A., Pentland, A., Lepri, B. & Letouzé, E. (2019) *Guide to mobile data analytics in refugee scenarios*. Cham: Springer.

Scheel, S. & Ustek-Spilda, F. (2018) *Big data, big promises: Revisiting migration statistics in context of the datafication of everything, Border Criminologies*. Available from: https://www.law.ox.ac.uk/research-subject-groups/centre-criminology/centreborder-criminologies/blog/2018/06/big-data-big [Accessed on 29 January 2022]

Sîrbu, A., Andrienko, G., Andrienko, N., Boldrini, C., Conti, M., Giannotti, F. et al. (2021) Human migration: the big data perspective. *International Journal of Data Science and Analytics*, 11(4), 341–360.

Verhulst, S.G. & Young, A. (2019) The potential and practice of data collaboratives for migration. In: Salah, A.A., Pentland, A., Lepri, B. & Letouzé, E. (Eds.) *Guide to mobile data analytics in refugee scenarios*. Cham: Springer, pp. 465–476.

How to cite this article: Salah, A.A. (2022) Can big data deliver its promises in migration research? *International Migration*, 60, 252–255. Available from: https://doi.org/10.1111/imig.12984